SURFACE COMPARISON WITH MASS TRANSPORTATION

Y. LIPMAN, I. DAUBECHIES

Abstract. We use mass-transportation as a tool to compare surfaces (2-manifolds). In particular, we determine the "similarity" of two given surfaces by solving a mass-transportation problem between their conformal densities. This mass transportation problem differs from the standard case in that we require the solution to be invariant under global Möbius transformations.
Our approach provides a constructive way of defining a metric in the abstract space of simply-connected smooth surfaces with boundary (i.e. surfaces of disk-type); this metric can also be used to define meaningful intrinsic distances between pairs of "patches" in the two surfaces, which allows automatic alignment of the surfaces. We provide numerical experiments on "real-life" surfaces to demonstrate possible applications in natural sciences.

1. introduction

Alignment of surfaces plays a role in a wide range of scientific disciplines. It is a standard problem in comparing different scans of manufactured objects; various algorithms have been proposed for this purpose in the computer graphics literature. It is often also a crucial step in a variety of problems in medicine and biology; in these cases the surfaces tend to be more complex, and the alignment problem may be harder. For instance, neuroscientists studying brain function through functional Magnetic Resonance Imaging (fMRI) typically observe several people performing identical tasks, obtaining readings for the corresponding activity in the brain cortex of each subject. In a first approximation, the cortex can be viewed as a highly convoluted 2-dimensional surface. Because different cortices are folded in very different ways, a synthesis of the observations from different subjects must be based on appropriate mappings between pairs of brain cortex surfaces, which reduces to a family of surface alignment problems [8, 27]. In another example, paleontologists studying molar teeth of mammals rely on detailed comparisons of the geometrical features of the tooth surfaces to distinguish species or to determine similarities or differences in diet [2].

Mathematically, the problem of surface alignment can be described as follows: given two 2-surfaces $\mathcal{M}$ and $\mathcal{N}$, find a mapping $f : \mathcal{M} \to \mathcal{N}$ that preserves, as best possible, "important properties" of the surfaces. The nature of the "important properties" depends on the problem at hand. In this paper, we concentrate on preserving the geometry, i.e., we would like the map f to preserve intrinsic distances, to the extent possible. In terms of the examples listed above, this is the criterion traditionally selected in the computer graphics literature; it also corresponds to the point of view of the paleontologists studying tooth surfaces. To align cortical surfaces, one typically uses the Talairach method [16] (which relies on geometrically defined landmarks and is thus geometric in nature as well), although alignment based on functional correspondences has been proposed more recently [27].

In this paper we propose a procedure to "geometrically" align surfaces, based on uniformization theory and optimal mass transportation. This approach is related to the computer graphics constructions in [18], which rely on the representation of isometries between topologically equivalent simply-connected surfaces by Möbius transformations between their uniformization spaces, and which exploit that 1) the Möbius group has small dimensionality (e.g. 3 for disk-type surfaces and 6 for sphere-type) and 2) changing the metric in one piece of a surface has little influence on the uniformization of distant parts. These two observations lead, in [18], to fast and particularly effective algorithms to identify near-isometries between differently deformed versions of a surface. In our present context, these same observations lead to a simple algorithm for surface alignment, reducing it to a linear programming problem.

We shall restrict ourselves to (sufficiently smooth) disk-type surfaces; we map them to metric densities defined on the hyperbolic disk, their canonical uniformization space. (Apart from simplifying the description of the surface, this also removes any effect of global translations and rotations on the description of each individual surface.) The alignment problem can then be studied in the framework of Kantorovich mass-transportation [14] between these metric densities, as follows. Mass-transportation seeks to minimize the "average distance" over which mass needs to be "moved" (in the most efficient such moving procedure) to transform one mass density μ into another, ν. In our case the uniformizing metric density (or conformal factor) corresponding to an initial surface is not unique, but is defined only up to a Möbius transformation. Because a naïve application of mass-transportation on the hyperbolic disk would not possess the requisite invariance under Möbius transformations, we generalize the mass-transportation framework, and replace the metric $d(x, y)$ traditionally used in defining the "average displacement distance" by a metric that depends on μ and ν, measuring the dissimilarity between the two metric densities on neighborhoods of x and y. Introducing neighborhoods also makes the definition less sensitive to noise in practical applications. The optimal way of transporting mass in this generalized framework, in which the orientation in space of the original surfaces is "factored away", automatically defines a corresponding optimal way of aligning the surfaces.

Our approach also allows us to define a new distance between surfaces. The average distance over which mass needs transporting (to transform one metric density into the other) quantifies the extent to which the two surfaces differ; we prove that it defines a distance metric between surfaces.

Other distances between surfaces have been used recently for several applications [19]. A prominent mathematical approach to define distances between surfaces considers the surfaces as special cases of *metric spaces*, and uses then the Gromov-Hausdorff (GH) distance between metric spaces [9]. The GH distance between metric spaces X and Y is defined through examining all the isometric embedding of X and Y into (other) metric spaces; although this distance possesses many attractive mathematical properties, it is inherently hard computationally [20, 1]. For instance, computing the GH distance is equivalent to a non-convex quadratic programming problem; solving this directly for correspondences is equivalent to integer quadratic assignment, and is thus NP-hard [5]. In addition, the non-convexity implies that the solution found in practice may be a local instead of a global minimum, and is therefore not guaranteed to give the correct answer for the GH distance. The distance metric between surfaces that we define in this paper does not have these shortcomings: because the computation of the distance between surfaces in our approach can be recast as a linear program, it can be implemented using efficient polynomial algorithms that are moreover guaranteed to converge to the correct solution.

It should be noted that in [19], Memoli generalizes the GH distance of [20] by introducing a quadratic mass transportation scheme to be applied to metric spaces already equipped with a measure (mm spaces); he notes that the computation of this Gromov-Wasserstein distance for mm spaces is somewhat easier and more stable to implement than the original GH distance. In our approach we do not need to equip the surfaces we compare with a measure: after uniformization reduces the problem to comparing two disks, we naturally "inherit" two corresponding conformal factors that we interpret as measure densities, for which we then apply an approach similar to the one proposed in [19]. Another crucial aspect in which our work differs from [19] is that, in contrast to the (continuous) quadratic programming method proposed in [19] to compute the Gromov-Wasserstein distance between mm spaces, our conformal approach leads to a convex (even linear) problem, solvable via a linear programming method.

It is worth mentioning that optimal mass transportation has been used as well, in the engineering literature to define interesting metrics between images; in this context metric is often called the Wasserstein distance. The seminal work for this image analysis approach is the paper by Rubner et al. [26], in which images are viewed as discrete measures, and the distance is called appropriately the "Earth Mover's Distance".

Another related method is presented in the papers of Zeng et al. [31, 32], which also use the uniformization space to match surfaces. Our work differs from that of Zeng et al. in that they use prescribed feature points (defined either by the user or by extra texture information) to calculate an interpolating harmonic map between the uniformization spaces, and then define the final correspondence as a composition of the

uniformization maps and this harmonic interpolant. This procedure is highly dependent on the prescribed feature points, provided as extra data or obtained from non-geometric information. In contrast, our work does not use any prescribed feature points, or external data, and makes use of only the geometry of the surface; in particular we make use of the conformal structure itself to define deviation from (local) isometry.

Our paper is organized as follows: in Section 2 we briefly recall some facts about uniformization and optimal mass transportation that we shall use, at the same time introducing our notation. Section 3 contains the main results of this paper, constructing the distance metric between disk-type surfaces, in several steps. Section 4 discusses various issues that concern the numerical implementation of the framework we propose; Section 5 illustrates our results with a few examples.

2. Background and Notations

As described in the introduction, our framework makes use of two mathematical theories: uniformization theory, to represent the surfaces as measures defined on a canonical domain, and optimal mass transportation, to align the measures. In this section we recall some of their basic properties, and we introduce our notations.

2.1. **Uniformization.** By the celebrated uniformization theory for Riemann surfaces (see for example [29, 11]), any simply-connected Riemann surface is conformally equivalent to one of three canonical domains: the sphere, the complex plane, or the unit disk. Since every 2-manifold surface $\mathcal{M}$ equipped with a smooth Riemannian metric g has an induced conformal structure and is thus a Riemann surface, uniformization applies to such surfaces. Therefore, every simply- connected surface with a Riemannian metric can be mapped conformally to one of the three canonical domains listed above. We shall consider surfaces $\mathcal{M}$ that are topologically equivalent to disks and that come equipped with a Riemannian metric tensor g (possibly inherited from the standard 3D metric if the surface is embedded in $\mathbb{R}^3$). For each such $\mathcal{M}$ there exists a conformal map $\phi : \mathcal{M} \to \mathcal{D}$, where $\mathcal{D} = \{z \mid |z| < 1\}$ is the open unit disk. The map ϕ pushes g to a metric on $\mathcal{D}$; denoting the coordinates in $\mathcal{D}$ by $z = x^1 + \mathbf{i}x^2$, we can write this metric as

$$\widetilde{g} = \phi_* g = \widetilde{\mu}(z)\, \delta_{ij}\, dx^i \otimes dx^j,$$

where $\widetilde{\mu}(z) > 0$, Einstein summation convention is used, and the subscript $*$ denotes the "push-forward" action. The function $\widetilde{\mu}$ can also be viewed as the *density function* of the measure $\mathrm{vol}_{\mathcal{M}}$ induced by the Riemann volume element: indeed, for (measurable) $A \subset \mathcal{M}$,

$$\mathrm{vol}_{\mathcal{M}}(A) = \int_{\phi(A)} \widetilde{\mu}(z)\, dx^1 \wedge dx^2. \tag{2.1}$$

It will be convenient to use the hyperbolic metric on the unit disk $(1 - |z|^2)^{-2}\delta_{ij}dx^i \otimes dx^j$ as a reference metric, rather than the standard Euclidean $\delta_{ij}dx^i \otimes dx^j$; note that they are conformally equivalent (with conformal factor $(1 - |z|^2)^{-2}$). Instead of the density $\widetilde{\mu}(z)$, we shall therefore use the *hyperbolic density function*

$$\mu^H(z) := (1 - |z|^2)^2\, \widetilde{\mu}(z)\,, \tag{2.2}$$

where the superscript H stands for hyperbolic. We shall often drop this superscript: unless otherwise stated $\mu = \mu^H$, and $\nu = \nu^H$ in what follows. The density function $\mu = \mu^H$ satisfies

$$\mathrm{vol}_{\mathcal{M}}(A) = \int_{\phi(A)} \mu(z)\, d\mathrm{vol}_H(z)\,,$$

where $d\mathrm{vol}_H(z) = (1 - |z|^2)^{-2}\, dx^1 \wedge dx^2$.

The conformal mappings of $\mathcal{D}$ to itself are the disk-preserving Möbius transformations $m \in M_D$, a family with three real parameters, defined by

$$m(z) = e^{\mathbf{i}\theta}\frac{z - a}{1 - \bar{a}z},\ a \in \mathcal{D},\ \theta \in [0, 2\pi). \tag{2.3}$$

Since these Möbius transformations satisfy

$$(1-|m(z)|^2)^{-2}|m'(z)|^2 = (1-|z|^2)^{-2}\,, \tag{2.4}$$

where m' stands for the derivatives of m, the pull-back of μ under a mapping $m \in M_D$ takes on a particularly simple expression. Setting $w = m(z)$, with $w = y^1 + \mathbf{i}y^2$, and $\widetilde{g}(w) = \widetilde{\mu}(w)\delta_{ij}dy^i \otimes dy^j = \mu(w)(1-|w|^2)^{-2}\delta_{ij}dy^i \otimes dy^j$, the definition

$$(m^*\widetilde{g})(z)_{kl}\, dx^k \otimes dx^\ell := \mu(w)\,(1-|w|^2)^{-2}\,\delta_{ij}\, dy^i \otimes dy^j$$

implies

$$\begin{aligned}(m^*\widetilde{g})_{k\ell}(z)\, dx^k \otimes dx^\ell &= \mu(m(z))(1-|m(z)|^2)^{-2}\,\delta_{ij}\,\frac{\partial y^i}{\partial x^k}\,\frac{\partial y^j}{\partial x^\ell}\, dx^k \otimes dx^\ell \\ &= \mu(m(z))\,(1-|m(z)|^2)^{-2}\,|m'(z)|^2\,\delta_{k\ell}\, dx^k \otimes dx^\ell \\ &= \mu(m(z))\,(1-|z|^2)^{-2}\,\delta_{k\ell}\, dx^k \otimes dx^\ell.\end{aligned}$$

In other words, $(m^*\widetilde{g})(z)_{kl}\, dx^k \otimes dx^\ell$ takes on the simple form $m^*\mu(z)\,(1-|z|^2)^{-2}\,\delta_{kl}\, dx^k \otimes dx^\ell$, with

$$m^*\mu(z) = \mu(m(z)). \tag{2.5}$$

Likewise, the push-forward, under a disk Möbius transform $m(z) = w$, of the diagonal Riemannian metric defined by the density function $\mu = \mu^H$, is again a diagonal metric, with (hyperbolic) density function $m_*\mu(w) = (m_*\mu)^H(w)$ given by

$$m_*\mu(w) = \mu(m^{-1}(w)). \tag{2.6}$$

It follows that checking whether or not two surfaces $\mathcal{M}$ and $\mathcal{N}$ are isometric, or searching for (near-) isometries between $\mathcal{M}$ and $\mathcal{N}$, is greatly simplified by considering the conformal mappings from $\mathcal{M}, \mathcal{N}$ to $\mathcal{D}$: once the (hyperbolic) density functions μ and ν are known, it suffices to identify $m \in M_D$ such that $\nu(m(z))$ equals $\mu(z)$ (or "nearly" equals, in a sense to be made precise). This was exploited in [18] to construct fast algorithms to find corresponding points between two given surfaces.

2.2. **Optimal mass transportation.** Optimal mass transportation was introduced by G. Monge [21], and L. Kantorovich [14]. It concerns the transformation of one mass distribution into another while minimizing a cost function that can be viewed as the amount of work required for the task. In the Kantorovich formulation, to which we shall stick in this paper, one considers two measure spaces X, Y, a probability measure on each, $\mu \in P(X)$, $\nu \in P(Y)$ (where $P(X), P(Y)$ are the respective probability measure spaces on X and Y), and the space $\Pi(\mu,\nu)$ of probability measures on $X \times Y$ with marginals μ and ν (resp.), that is, for $A \subset X$, $B \subset Y$, $\pi(A \times Y) = \mu(A)$ and $\pi(X \times B) = \nu(B)$. The *optimal* mass transportation is the element of $\Pi(\mu,\nu)$ that minimizes $\int_{X\times Y} d(x,y)d\pi(x,y)$, where $d(x,y)$ is a cost function. (In general, one should consider an infimum rather than a minimum; in our case, X and Y are compact, $d(\cdot,\cdot)$ is continuous, and the infimum is achieved.) The corresponding minimum,

$$T_d^R(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} d(x,y)d\pi(x,y), \tag{2.7}$$

is the optimal mass transportation distance between μ and ν, with respect to the cost function $d(x,y)$.

Intuitively, one can interpret this as follows: imagine being confronted with a pile of sand on the one hand (μ), and a hole in the ground on the other hand ($-\nu$), and assume that the volume of the sand pile equals exactly the volume of the hole (suitably normalized, μ, ν are probability measures). You wish to fill the hole with the sand from the pile ($\pi \in \Pi(\mu,\nu)$), in a way that minimizes the amount of work (represented by $\int d(x,y)d\pi(x,y)$, where $d(\cdot,\cdot)$ can be thought of as a distance function). In the engineering literature, the distance $T_d^R(\mu,\nu)$ is often called the "earth mover's distance" [26], a name that echoes this intuition.

In what follows, we shall apply this framework to the density functions μ and ν on the hyperbolic disk $\mathcal{D}$ obtained by conformal mappings from two surfaces $\mathcal{M}, \mathcal{N}$, as described in the previous subsection.

The main obstacle to applying the Kantorovich transportation framework directly is that the density μ, characterizing the Riemannian metric on $\mathcal{D}$ obtained by pushing forward the metric on $\mathcal{M}$ via the uniformizing map $\phi : \mathcal{M} \to \mathcal{D}$, is not uniquely defined: another uniformizing map $\phi' : \mathcal{M} \to \mathcal{D}$ may well produce a different μ'. Because the two representations are necessarily isometric ($\phi^{-1} \circ \phi'$ maps $\mathcal{M}$ isometrically to itself), we must have $\mu'(m(z)) = \mu(z)$ for some $m \in M_D$. (In fact, $m = \phi' \circ \phi^{-1}$.) In a sense, the representation of (disk-type) surfaces $\mathcal{M}$ as measures over $\mathcal{D}$ should be considered "modulo" the disk Möbius transformations.

We thus need to address how to adapt the optimal transportation framework to factor out this Möbius transformation ambiguity. This is done by designing a special distance (or cost) functional $d^R_{\mu,\nu}(z, w)$ that *depends* on the conformal densities μ and ν representing the two surfaces. (A fairly simple argument shows that a cost function that does not depend on μ and ν allows only trivial answers, such as $d(z, w) = 0$ for all z, w.) As we shall see in the next section, this cost function will have an intuitive explanation: $d^R_{\mu,\nu}(z, w)$ will measure how well an R-sized neighborhood of z with density μ can be matched isometrically to an R-sized neighborhood of w with density ν by means of a disk Möbius transformation.

3. Optimal volume transportation for surfaces

We want to measure distances between surfaces by using the Kantorovich transportation framework to measure the transportation between the metric densities on $\mathcal{D}$ obtained by uniformization applied to the surfaces. The main obstacle is that these metric densities are not uniquely defined; they are defined up to a Möbius transformation. In particular, if two densities μ and ν are related by $\nu = m_*\mu$ (i.e. $\mu(z) = \nu(m(z))$), where $m \in M_D$, then we want our putative distance between μ and ν to be zero, since they describe isometric surfaces, and could have been obtained by different uniformization maps of the same surface. A standard approach to obtain quantities that are invariant under the operation of some group (in our case, the disk Möbius transformations) is by minimizing over the possible group operations. For instance, we could set

$$\text{Distance}(\mu, \nu) = \inf_{m \in M_D} \left(\inf_{\pi \in \Pi(m_*\mu, \nu)} \int_{\mathcal{D}\times\mathcal{D}} d(z, w)\, d\pi(z, w) \right),$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathcal{D} \times \mathcal{D}$ with marginals $\mu \,\mathrm{vol}_H$ and $\nu \,\mathrm{vol}_H$. In order for this to be computationally feasible, we would want the minimum to be achieved in some m, which would depend on μ and ν of course; let's denote this special minimizing $m \in M_D$ by $m_{\mu,\nu}$. This would mean

$$\begin{aligned} \text{Distance}(\mu, \nu) &= \inf_{\pi \in \Pi([m_{\mu,\nu}]_*\mu, \nu)} \int_{\mathcal{D}\times\mathcal{D}} d(z, w)\, d\pi(z, w) \\ &= \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{D}\times\mathcal{D}} d(m_{\mu,\nu}(z), w)\, d\pi(z, w)\,. \end{aligned} \tag{3.1}$$

If ν were itself already equal to $m'_*\mu$, for some $m' \in M_D$, then we would expect the minimizing Möbius transformation to be $m_{\mu,\nu} = m'$; for π supported on the diagonal $\mathtt{d} = \{(z, z)\,;\, z \in \mathcal{D}\} \subset \mathcal{D} \times \mathcal{D}$, defined by $\pi(A) = \int_{A_2} \nu(w)\, d\mathrm{vol}_H(w)$, with $A_2 = \{w; (w, w) \in A\}$, one would then indeed have $\int_{\mathcal{D}\times\mathcal{D}} d(m_{\mu,\nu}(z), w)\, d\pi(z, w) = 0$, leading to $\text{Distance}(\mu, m'_*\mu) = 0$. From (3.1) one sees that this amounts to using the same formula as for the standard Kantorovich approach with just one change: *the cost function depends on* μ and ν.

We shall use a variant on this construction, retaining the principle of using cost functions $d(\cdot, \cdot)$ in the integrand that depend on μ and ν, without picking them necessarily of the form $d(m_{\mu,\nu}(z), w)$. In addition to introducing such a dependence, we also wish to incorporate some robustness into the evaluation of the distance between (or dissimilarity of) μ and ν. We shall do this by using a cost function $d^R_{\mu,\nu}(z, w)$ that depends on a comparison of the behavior μ and ν on *neighborhoods* of z and w, mapped by m ranging over M_D. The next subsection shows precisely how this is done.

3.1. Construction of $d^R_{\mu,\nu}(z, w)$.

We construct $d^R_{\mu,\nu}(z, w)$ so that it indicates the extent to which a neighborhood of the point z in $(\mathcal{D}, \mu)$, the (conformal representation of the) first surface, is isometric with a neighborhood of the point w in $(\mathcal{D}, \nu)$, the (conformal representation of the) second surface. We will need to define

two ingredients for this: the neighborhoods we will use, and how we shall characterize the (dis)similarity of two neighborhoods, equipped with different metrics.

We start with the neighborhoods.

For a fixed radius $R > 0$, we define $\Omega_{z_0,R}$ to be the hyperbolic geodesic disk of radius R centered at z_0. The following gives an easy procedure to construct these disks. If $z_0 = 0$, then the hyperbolic geodesic disks centered at $z_0 = 0$ are also "standard" (i.e. Euclidean) disks centered at 0: $\Omega_{0,R} = \{z\,;\,|z| \le r_R\}$, where $r_R = \operatorname{arctanh}(r) = R$. The hyperbolic disks around other centers are images of these central disks under Möbius transformations (= hyperbolic isometries): setting $m(z) = (z - z_0)(1 - z\bar{z_0})^{-1}$, we have

$$\Omega_{z_0,R} = m^{-1}(\Omega_{0,R})\,. \tag{3.2}$$

If m', m'' are two maps in M_D that both map z_0 to 0, then $m'' \circ (m')^{-1}$ simply rotates $\Omega_{0,R}$ around its center, over some angle θ determined by m' and m''. From this observation one easily checks that (3.2) holds for *any* $m \in M_D$ that maps z_0 to 0. In fact, we have the following more general

Lemma 3.1. *For arbitrary $z, w \in \mathcal{D}$ and any $R > 0$, every disk Möbius transformation $m \in M_D$ that maps z to w (i.e. $w = m(z)$) also maps $\Omega_{z,R}$ to $\Omega_{w,R}$.*

Next we define how to quantify the (dis)similarity of the pairs $(\Omega_{z_0,R}\,,\,\mu\,)$ and $(\Omega_{w_0,R}\,,\,\nu\,)$. Since (global) isometries are given by the elements of the disk-preserving Möbius group M_D, we will test the extent to which the two patches are isometric by comparing $(\Omega_{w_0,R}\,,\,\nu\,)$ with all the images of $(\Omega_{z_0,R}\,,\,\mu\,)$ under Möbius transformations in M_D that take z_0 to w_0.

To carry out this comparison, we need a norm. Any metric $g_{ij}(z)dx^i \otimes dx^j$ induces an inner product on the space of 2-covariant tensors, as follows: if $\mathbf{a}(z) = a_{ij}(z)\,dx^i \otimes dx^j$ and $\mathbf{b}(z) = b_{ij}(z)\,dx^i \otimes dx^j$ are two 2-covariant tensors in our parameter space $\mathcal{D}$, then their inner product is defined by

$$\langle \mathbf{a}(z), \mathbf{b}(z) \rangle = a_{ij}(z)\,b_{k\ell}(z)\,g^{ik}(z)\,g^{j\ell}(z)\;; \tag{3.3}$$

as always, this inner product defines a norm, $\|\mathbf{a}\|_z^2 = a_{ij}(z)\,a_{k\ell}(z)\,g^{ik}(z)\,g^{j\ell}(z)$.

Now, let us apply this to the computation of the norm of the difference between the local metric on one surface, $g_{ij}(z) = \mu(z)(1 - |z|^2)^{-2}\delta_{ij}$, and $h_{ij}(w) = \nu(w)(1 - |w|^2)^{-2}\delta_{ij}$, the pull-back metric from the other surface by a Möbius transformation m. Using (3.3),(2.5), and writing $\boldsymbol{\delta}$ for the tensor with entries δ_{ij}, we have:

$$\begin{aligned}\|\mu - m^*\nu\|_z^2 &= \|\,\mu(z)(1 - |z|^2)^{-2}\boldsymbol{\delta} - \nu(m(z))(1 - |z|^2)^{-2}\boldsymbol{\delta}\,\|_z^2 \\ &= \Big(\mu(z) - \nu(m(z))\Big)^2(1 - |z|^2)^{-4}\,\delta_{ij}\,\delta_{k\ell}\,g^{ik}(z)\,g^{j\ell}(z) = \left(1 - \frac{\nu(m(z))}{\mu(z)}\right)^2.\end{aligned}$$

We are now ready to define the distance function $d^R_{\mu,\nu}(z, w)$:

$$d^R_{\mu,\nu}(z_0, w_0) := \inf_{m \in M_D\,,\, m(z_0) = w_0} \int_{\Omega_{z_0,R}} |\,\mu(z) - (m^*\nu)(z)\,|\,d\mathrm{vol}_H(z), \tag{3.4}$$

where $d\mathrm{vol}_H(z) = (1 - |z|^2)^{-2}\,dx \wedge dy$ is the volume form for the hyperbolic disk. The integral in (3.4) can also be written in the following form, which makes its invariance more readily apparent:

$$\int_{\Omega_{z_0,R}} \left|1 - \frac{\nu(m(z))}{\mu(z)}\right| d\mathrm{vol}_{\mathcal{M}}(z) = \int_{\Omega_{z_0,R}} \|\mu - m^*\nu\|_z\,d\mathrm{vol}_{\mathcal{M}}(z), \tag{3.5}$$

where $d\mathrm{vol}_{\mathcal{M}}(z) = \mu(z)(1 - |z|^2)^{-2}\,dx^1 \wedge dx^2 = \sqrt{|g_{ij}|}\,dx^1 \wedge dx^2$ is the volume form of the first surface $\mathcal{M}$.

The next Lemma shows that although the integration in (3.5) is carried out w.r.t. the volume of the first surface, this measure of distance is nevertheless symmetric:

Lemma 3.2. *If $m \in M_D$ maps z_0 to w_0, $m(z_0) = w_0$, then*

$$\int_{\Omega_{z_0,R}} \Big| \mu(z) - m^*\nu(z) \Big| \, d\mathrm{vol}_H(z) = \int_{\Omega_{w_0,R}} \Big| m_*\mu(w) - \nu(w) \Big| \, d\mathrm{vol}_H(w).$$

Proof. By the pull-back formula (2.5), we have

$$\int_{\Omega_{z_0,R}} \Big| \mu(z) - m^*\nu(z) \Big| \, d\mathrm{vol}_H(z) = \int_{\Omega_{z_0}} \Big| \mu(z) - \nu(m(z)) \Big| \, d\mathrm{vol}_H(z).$$

Performing the change of coordinates $z = m^{-1}(w)$ in the integral on the right hand side, we obtain

$$\int_{m(\Omega_{z_0,R})} \Big| \mu(m^{-1}(w)) - \nu(w) \Big| \, d\mathrm{vol}_H(w),$$

where we have used that m^{-1} is an isometry and therefore preserves the volume element $d\mathrm{vol}_H(w) = (1 - |w|^2)^{-2}\, dy^1 \wedge dy^2$. By Lemma 3.1, $m(\Omega_{z_0,R}) = \Omega_{w_0,R}$; using the push-forward formula (2.6) then allows to conclude. □

Note that our point of view in defining our "distance" between z and w differs from the classical point of view in mass transportation: Traditionally, $d(z, w)$ is some sort of *physical distance* between the points z and w; in our case $d^R_{\mu,\nu}(z, w)$ measures the dissimilarity of (neighborhoods of) z and w.

The next Theorem lists some important properties of $d^R_{\mu,\nu}$; its proof is given in Appendix A.

Theorem 3.3. *The distance function $d^R_{\mu,\nu}(z, w)$ satisfies the following properties*

(1)	$d^R_{m_1^*\mu,m_2^*\nu}(m_1^{-1}(z_0), m_2^{-1}(w_0)) = d^R_{\mu,\nu}(z_0, w_0)$	Invariance under (well-defined) Möbius changes of coordinates
(2)	$d^R_{\mu,\nu}(z_0, w_0) = d^R_{\nu,\mu}(w_0, z_0)$	Symmetry
(3)	$d^R_{\mu,\nu}(z_0, w_0) \geq 0$	Non-negativity
(4)	$d^R_{\mu,\nu}(z_0, w_0) = 0 \implies \Omega_{z_0,R}$ in $(\mathcal{D}, \mu)$ and $\Omega_{w_0,R}$ in $(\mathcal{D}, \nu)$ are isometric	
(5)	$d^R_{m^*\nu,\nu}(m^{-1}(z_0), z_0) = 0$	Reflexivity
(6)	$d^R_{\mu_1,\mu_3}(z_1, z_3) \leq d^R_{\mu_1,\mu_2}(z_1, z_2) + d^R_{\mu_2,\mu_3}(z_2, z_3)$	Triangle inequality

In addition, the function $d^R_{\mu,\nu} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is continuous. To show this, we first look a little more closely at the family of disk Möbius transformations that map one pre-assigned point $z_0 \in \mathcal{D}$ to another pre-assigned point $w_0 \in \mathcal{D}$, over which one minimizes to define $d^R_\mu(z_0, w_0)$.

Definition 3.4. For any pair of points $z_0,\ w_0 \in \mathcal{D}$, we denote by M_{D,z_0,w_0} the set of Möbius transformations that map z_0 to w_0.

This family of Möbius transformations is completely characterized by the following lemma:

Lemma 3.5. *For any $z_0, w_0 \in \mathcal{D}$, the set M_{D,z_0,w_0} constitutes a 1-parameter family of disk Möbius transformations, parametrized continuously over S^1 (the unit circle). More precisely, every $m \in M_{D,z_0,w_0}$ is of the form*

$$m(z) = \tau\, \frac{z - a}{1 - \overline{a} z}\ , \quad \text{with} \quad a = a(z_0, w_0, \sigma) := \frac{z_0 - w_0\, \overline{\sigma}}{1 - \overline{z_0}\, w_0\, \overline{\sigma}} \ \text{ and } \ \tau = \tau(z_0, w_0, \sigma) := \sigma \frac{1 - \overline{z_0}\, w_0\, \overline{\sigma}}{1 - z_0\, \overline{w_0}\, \sigma}, \tag{3.6}$$

where $\sigma \in S_1 := \{z \in \mathbb{C}\,;\, |z| = 1\}$ can be chosen freely.

Proof. By (2.3), the disk Möbius transformations that map z_0 to 0 all have the form

$$m_{\psi,z_0}(z) = e^{\mathbf{i}\psi}\,\frac{z-z_0}{1-\overline{z_0}\,z}\,, \quad \text{the inverse of which is } \; m^{-1}_{\psi,z_0}(w) = e^{-\mathbf{i}\psi}\,\frac{w+e^{\mathbf{i}\psi}z_0}{1+e^{-\mathbf{i}\psi}\,\overline{z_0}w}\;,$$

where $\psi \in \mathbb{R}$ can be set arbitrarily. It follows that the elements of M_{D,z_0,w_0} are given by the family $m^{-1}_{\gamma,w_0} \circ m_{\psi,z_0}$, with ψ, $\gamma \in \mathbb{R}$. Working this out, one finds that these combinations of Möbius transformations take the form (3.6), with $\sigma = e^{\mathbf{i}(\psi-\gamma)}$. □

We shall denote by $m_{z_0,w_0,\sigma}$ the special disk Möbius transformation defined by (3.6). In view of our interest in $d^R_{\mu,\nu}$, we also define the auxiliary function

$$\Phi : \mathcal{D}\times\mathcal{D}\times S_1 \longrightarrow \mathbb{C}$$

by $\Phi(z_0,w_0,\sigma) = \int_{\Omega(z_0,R)} |\,\mu(z) - \nu(m_{z_0,w_0,\sigma}(z))\,|\,d\mathrm{vol}_H(z)$. This function has the following continuity properties, inherited from μ and ν:

Lemma 3.6.
- *For each fixed (z_0,w_0), the function $\Phi(z_0,w_0,\cdot)$ is continuous on S_1.*
- *For each fixed $\sigma \in S_1$, $\Phi(\cdot,\cdot,\sigma)$ is continuous on $\mathcal{D}\times\mathcal{D}$. Moreover, the family $\Big(\Phi(\cdot,\cdot,\sigma)\Big)_{\sigma\in S_1}$ is equicontinuous.*

Proof. The proof of this Lemma is given in Appendix A. □

Note that since S^1 is compact, Lemma 3.6 implies that the infimum in the definition of $d^R_{\mu,\nu}$ can be replaced by a minimum:

$$d^R_{\mu,\nu}(z_0,w_0) = \min_{m(z_0)=w_0} \int_{\Omega_{z_0,R}} |\,\mu(z) - \nu(m(z))\,|\,d\mathrm{vol}_H(z)\;.$$

We have now all the building blocks to prove

Theorem 3.7. *If μ and ν are continuous from $\mathcal{D}$ to $\mathbb{R}$, then $d^R_{\mu,\nu}(z,w)$ is a continuous function on $\mathcal{D}\times\mathcal{D}$.*

Proof. Pick an arbitrary point $(z_0,w_0) \in \mathcal{D}\times\mathcal{D}$, and pick $\varepsilon > 0$ arbitrarily small.

By Lemma 3.6, there exists a $\delta > 0$ such that, for $|z_0' - z_0| < \delta$, $|w_0' - w_0| < \delta$, we have

$$|\,\Phi(z_0,w_0,\sigma) - \Phi(z_0',w_0',\sigma)\,| \;\le\; \varepsilon\;,$$

uniformly in σ. Pick now arbitrary z_0', w_0' so that $|z_0 - z_0'|, |w_0 - w_0'| < \delta$.

Let $m_{z_0,w_0,\sigma}$, resp. $m_{z_0',w_0',\sigma'}$, be the minimizing Möbius transform in the definition of $d^R_{\mu,\nu}(z_0,w_0)$, resp. $d^R_{\mu,\nu}(z_0',w_0')$, i.e.

$$d^R_{\mu,\nu}(z_0,w_0) = \Phi(z_0,w_0,\sigma) \;\text{ and }\; d^R_{\mu,\nu}(z_0',w_0') = \Phi(z_0,w_0,\sigma')\;.$$

It then follows that

$$\begin{aligned} d^R_{\mu,\nu}(z_0,w_0) &= \min_\tau \Phi(z_0,w_0,\tau) \le \Phi(z_0,w_0,\sigma') \\ &\le \Phi(z_0',w_0',\sigma') + |\Phi(z_0,w_0,\sigma') - \Phi(z_0',w_0',\sigma')| = d^R_{\mu,\nu}(z_0',w_0') + |\Phi(z_0,w_0,\sigma') - \Phi(z_0',w_0',\sigma')| \\ &\le d^R_{\mu,\nu}(z_0',w_0') + \sup_{\omega\in S_1} |\Phi(z_0,w_0,\omega) - \Phi(z_0',w_0',\omega)| \le d^R_{\mu,\nu}(z_0',w_0') + \varepsilon\;. \end{aligned}$$

Likewise $d^R_{\mu,\nu}(z_0',w_0') \le d^R_{\mu,\nu}(z_0,w_0) + \varepsilon$, so that $\big|d^R_{\mu,\nu}(z_0,w_0) - d^R_{\mu,\nu}(z_0',w_0')\big| < \varepsilon$. □

3.2. **Incorporating $d^R_{\mu,\nu}(z,w)$ into the transportation framework.** The next step in constructing the distance operator between surfaces is to incorporate the distance $d^R_{\mu,\nu}(z,w)$ defined in the previous subsection into the (generalized) Kantorovich transportation model:

$$T^R_d(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w)d\pi(z,w). \tag{3.7}$$

The main result is that this procedure (under some extra conditions) furnishes a *metric* between (disk-type) surfaces.

Theorem 3.8. *There exists $\pi^* \in \Pi(\mu,\nu)$ such that*

$$\int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w)d\pi^*(z,w) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w)d\pi(z,w).$$

Proof. This proof follows the same argument as in [30], adapted here to our generalized setting. It uses the continuity of the distance function to derive the existence of a global minimum of (3.7). Let $\left(\pi_k\right)_{k\in\mathbb{N}} \in \Pi(\mu,\nu)$ be a minimizer sequence of (3.7), for example by taking

$$\int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w)d\pi_k(z,w) < T^R_d(\mu,\nu) + \frac{1}{k}.$$

Then this sequence of measures is tight, that is, for every $\varepsilon > 0$, there exists a compact set $C \subset \mathcal{D}\times\mathcal{D}$ such that $\pi_k(C) > 1-\varepsilon$, for all $k \in \mathbb{N}$. To see this, note that since $\mathcal{D}$ is separable and complete, the measures μ, ν are *tight* measures (see [22]). This means that for arbitrary $\varepsilon > 0$, there exist compact sets $A, B \subset \mathcal{D}$ so that $\mu(A) > 1-\varepsilon/2$ and $\nu(B) > 1-\varepsilon/2$. It then follows that, for all $k \in \mathbb{N}$,

$$\pi_k(A\times B) = \pi_k(A\times\mathcal{D}) - \pi_k(A\times(\mathcal{D}\setminus B)\,) \geq \mu(A) - \nu(\mathcal{D}\setminus B) = \mu(A) - (1-\nu(B)) > 1-\varepsilon.$$

Since the set $C = A\times B \subset \mathcal{D}\times\mathcal{D}$ is compact, this proves the claimed tightness of the family $\left(\pi_k\right)_{k\in\mathbb{N}}$. By Prohorov's Theorem [22], a tight family of measures is sequentially weakly compact; in our case this means that $\left(\pi_k\right)_{k\in\mathbb{N}}$ has a weakly convergent subsequence $\left(\pi_{k_n}\right)_{n\in\mathbb{N}}$; by definition, its weak limit π^* satisfies, for every bounded continuous function f on $\mathcal{D}\times\mathcal{D}$,

$$\int_{\mathcal{D}\times\mathcal{D}} f(z,w)d\pi_{k_n}(z,w) \to \int_{\mathcal{D}\times\mathcal{D}} f(z,w)d\pi^*(z,w).$$

Therefore, taking in particular the continuous function $f(z,w) = d^R_{\mu,\nu}(z,w)$, we obtain

$$T^R_d(\mu,\nu) = \lim_{n\to\infty} \int_{\mathcal{D}\times\mathcal{D}} f(z,w)d\pi_{k_n}(z,w) = \int_{\mathcal{D}\times\mathcal{D}} f(z,w)d\pi^*(z,w).$$

□

Under rather mild conditions, the "standard" Kantorovich transportation (2.7) on a metric spaces (X,d) defines a metric on the space of probability measures on X . We will prove that our generalization defines a distance metric as well. More precisely, we shall prove first that

$$\mathbf{d}^R(\mathcal{M},\mathcal{N}) = T^R_d(\mu,\nu)$$

defines a semi-metric in the set of all disk-type surfaces. We shall restrict ourselves to surfaces that are sufficiently smooth to allow uniformization, so that they can be globally and conformally parameterized over the hyperbolic disk. Under some extra assumptions, we will prove that $\mathbf{d}^R$ is a metric, in the sense that $\mathbf{d}^R(\mathcal{M},\mathcal{N}) = 0$ implies that $\mathcal{M}$ and $\mathcal{N}$ are isometric.

For the semi-metric part we will again adapt a proof given in [30] to our framework. In particular, we shall make use of the following "gluing lemma":

Lemma 3.9. *Let μ_1, μ_2, μ_3 be three probability measures on $\mathcal{D}$, and let $\pi_{12} \in \Pi(\mu_1, \mu_2)$, $\pi_{23} \in \Pi(\mu_2, \mu_3)$ be two transportation plans. Then there exist a probability measure π on $\mathcal{D} \times \mathcal{D} \times \mathcal{D}$ that has π_{12}, π_{23} as marginals, that is $\int_{z_3 \in \mathcal{D}} d\pi(z_1, z_2, z_3) = d\pi_{12}(z_1, z_2)$, and $\int_{z_1 \in \mathcal{D}} d\pi(z_1, z_2, z_3) = d\pi_{23}(z_2, z_3)$.*

This lemma will be used in the proof of the following:

Theorem 3.10. *For two disk-type surfaces $\mathcal{M} = (\mathcal{D}, \mu)$, $\mathcal{N} = (\mathcal{D}, \nu)$, let $\mathbf{d}^R(\mathcal{M}, \mathcal{N})$ be defined by*

$$\mathbf{d}^R(\mathcal{M}, \mathcal{N}) = T_d^R(\mu, \nu).$$

Then $\mathbf{d}^R$ defines a semi-metric on the space of disk-type surfaces.

Proof. The symmetry of $d^R_{\mu,\nu}$ implies symmetry for T_d^R, by the following argument:

$$\begin{aligned} T_d^R(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z, w) d\pi(z, w) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{D}\times\mathcal{D}} d^R_{\nu,\mu}(w, z) d\pi(z, w) \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{D}\times\mathcal{D}} d^R_{\nu,\mu}(w, z) d\widetilde{\pi}(w, z), \quad \text{where we have set } \ \widetilde{\pi}(w, z) = \pi(z, w) \\ &= T_d^R(\nu, \mu) \ . \ (\text{ use that } \pi \in \Pi(\mu, \nu) \Leftrightarrow \widetilde{\pi} \in \Pi(\nu, \mu)) \end{aligned}$$

The non-negativity of $d^R_{\mu,\nu}(\cdot, \cdot)$ automatically implies $T_d^R(\mu, \nu) \geq 0$.

Next we show that, for any Möbius transformation m, $T_d^R(\mu, m_*\mu) = 0$. To see this, pick the transportation plan $\pi \in \Pi(\mu, m_*\mu)$ defined by

$$\int_{\mathcal{D}\times\mathcal{D}} f(z, w) d\pi(z, w) = \int_{\mathcal{D}} f(z, m(z))\mu(z)\, d\mathrm{vol}_H(z).$$

On the one hand $\pi \in \Pi(\mu, m_*\mu)$, since

$$\int_{A\times\mathcal{D}} d\pi(z, w) = \int_A \mu(z) d\mathrm{vol}_H(z),$$

and

$$\begin{aligned} \int_{\mathcal{D}\times B} d\pi(z, w) &= \int_{\mathcal{D}\times\mathcal{D}} \chi_B(w) d\pi(z, w) \\ &= \int_{\mathcal{D}} \chi_B(m(z))\mu(z) d\mathrm{vol}_H(z) = \int_{\mathcal{D}} \chi_B(w)\mu_*(w) d\mathrm{vol}_H(w), \end{aligned}$$

where we used the change of variables $w = m(z)$ in the last step. Furthermore, $\pi(z, w)$ is concentrated on the graph of m, i.e. on $\{(z, m(z)) \ ; \ z \in \mathcal{D}\} \subset \mathcal{D} \times \mathcal{D}$. Since $d^R_{\mu,m_*\mu}(z, m(z)) = 0$ for all $z \in \mathcal{D}$ we obtain therefore $T_d(\mu, m_*\mu) \leq \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,m_*\mu}(z, w) d\pi(z, w) = 0$.

Finally, we prove the triangle inequality $T_d^R(\mu_1, \mu_3) \leq T_d^R(\mu_1, \mu_2) + T_d^R(\mu_2, \mu_3)$. To this end we follow the argument in the proof given in [30] (page 208). This is where we invoke the gluing Lemma stated above.

We start by picking arbitrary transportation plans $\pi_{12} \in \Pi(\mu_1, \mu_2)$ and $\pi_{23} \in \Pi(\mu_2, \mu_3)$. By Lemma 3.9 there exists a probability measure π on $\mathcal{D} \times \mathcal{D} \times \mathcal{D}$ with marginals π_{12} and π_{23}. Denote by π_{13} its third marginal, that is

$$\int_{z_2 \in \mathcal{D}} d\pi(z_1, z_2, z_3) = d\pi_{13}(z_1, z_3).$$

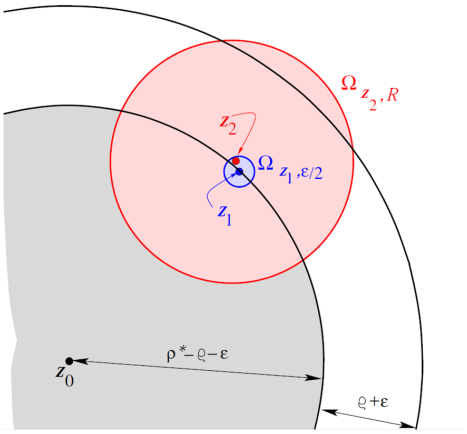


FIGURE 1. Illustration of the proof of Theorem 3.14

Then

$$\begin{aligned} T_d^R(\mu_1,\mu_3) &\le \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu_1,\mu_3}(z_1,z_3)d\pi_{13}(z_1,z_3) = \int_{\mathcal{D}\times\mathcal{D}\times\mathcal{D}} d^R_{\mu_1,\mu_3}(z_1,z_3)d\pi(z_1,z_2,z_3) \\ &\le \int_{\mathcal{D}\times\mathcal{D}\times\mathcal{D}} \left(d^R_{\mu_1,\mu_2}(z_1,z_2) + d^R_{\mu_2,\mu_3}(z_2,z_3)\right)d\pi(z_1,z_2,z_3) \\ &\le \int_{\mathcal{D}\times\mathcal{D}\times\mathcal{D}} d^R_{\mu_1,\mu_2}(z_1,z_2)d\pi(z_1,z_2,z_3) + \int_{\mathcal{D}\times\mathcal{D}\times\mathcal{D}} d^R_{\mu_2,\mu_3}(z_2,z_3)d\pi(z_1,z_2,z_3) \\ &\le \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu_1,\mu_2}(z_1,z_2)d\pi_{12}(z_1,z_2) + \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu_2,\mu_3}(z_2,z_3)d\pi_{23}(z_2,z_3), \end{aligned}$$

where we used the triangle-inequality for $d^R_{\mu,\nu}$ listed in (Theorem 3.3). Since we can choose π_{12} and π_{23} to achieve arbitrary close values to the infimum in eq. (3.7) the triangle inequality follows. □

To qualify as a metric rather than a semi-metric, $\mathbf{d}^R$ (or T_d^R) should be able to distinguish from each other any two surfaces (or measures) that are not "identical", that is isometric. To prove that they can do so, we need an extra assumption: we shall require that the surfaces we consider have no self-isometries. More precisely, we require that each surface $\mathcal{M}$ that we consider satisfies the following definition:

Definition 3.11. A surface $\mathcal{M}$ is said to be a singly ϱ-$_{\mathrm{H}}$fittable (where $\varrho \in \mathbb{R}$, $\varrho \neq 0$) if, for all $R > \varrho$, and all $z \in \mathcal{D}$, there is no other Möbius transformation m other than the identity for which

$$\int_{\Omega_{z,R}} |\mu(z) - \mu(m(z))|\, d\mathrm{vol}_H(z) = 0.$$

Remark 3.12. This definition can also be read as follows: $\mathcal{M}$ is singly ϱ-$_{\mathrm{H}}$fittable if and only if, for all $R > \varrho$, any two conformal factors μ_1 and μ_2 for $\mathcal{M}$ satisfy:

(1) For all $z \in \mathcal{D}$ there exists a unique minimum to the function $w \mapsto d^R_{\mu_1,\mu_2}(z,w)$.
(2) For all pairs $(z,w) \in \mathcal{D}\times\mathcal{D}$ that achieve this minimum there exists a unique Möbius transformation for which the integral in (3.4) vanishes (with μ_1 in the role of μ, and μ_2 in that of ν).

Essentially, this definition requires that, from some sufficiently large (hyperbolic) scale onwards, there are no isometric pieces within $(\mathcal{D},\mu)$ (or $(\mathcal{D},\nu)$).

We start with a lemma, and then prove the main result of this subsection.

Lemma 3.13. *Let $\pi \in \Pi(\mu,\nu)$ be such that $\int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w)\, d\pi(z,w) = 0$. Then, for all $z_0 \in \mathcal{D}$ and $\delta > 0$, there exists at least one point $z \in \Omega_{z_0,\delta}$ such that $d^R_{\mu,\nu}(z,w) = 0$ for some $w \in \mathcal{D}$.*

Proof. By contradiction: assume that there exists a disk $\Omega_{z_0,\delta}$ such that $d^R_{\mu,\nu}(z,w) > 0$ for all $z \in \Omega_{z_0,\delta}$ and all $w \in \mathcal{D}$. Since

$$\int_{\Omega(z_0,\delta)\times\mathcal{D}} d\pi(z,w) = \int_{\Omega(z_0,\delta)} \mu(z)\, d\mathrm{vol}_H(z) > 0 \ ,$$

the set $\Omega(z_0,\delta)\times\mathcal{D}$ contains some of the support of π. It follows that

$$\int_{\Omega(z_0,\delta)\times\mathcal{D}} d^R_{\mu,\nu}(z,w) d\pi(z,w) > 0 \ ,$$

which contradicts

$$\int_{\Omega(z_0,\delta)\times\mathcal{D}} d^R_{\mu,\nu}(z,w) d\pi(z,w) \le \int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w) d\pi(z,w) = 0 \ .$$

□

Theorem 3.14. *Suppose that $\mathcal{M}$ and $\mathcal{N}$ are two surfaces that are singly ϱ-$_H$fittable. If $\mathbf{d}^R(\mathcal{M},\mathcal{N}) = 0$ for some $R > \varrho$, then there exists a Möbius transformation $m \in M_D$ that is a global isometry between $\mathcal{M} = (\mathcal{D},\mu)$ and $\mathcal{N} = (\mathcal{D},\nu)$ (where μ and ν are conformal factors of $\mathcal{M}$ and $\mathcal{N}$, respectively).*

Proof. When $\mathbf{d}^R(\mathcal{M},\mathcal{N}) = 0$, there exists (see [30]) $\pi \in \Pi(\mu,\nu)$ such that

$$\int_{\mathcal{D}\times\mathcal{D}} d^R_{\mu,\nu}(z,w) d\pi(z,w) = 0.$$

Next, pick an arbitrary point $z_0 \in \mathcal{D}$ such that, for some $w_0 \in \mathcal{D}$, we have $d^R_{\mu,\nu}(z_0,w_0) = 0$. (The existence of such a pair is guaranteed by Lemma 3.13.) This implies that there exists a unique Möbius transformation $m_0 \in M_D$ that takes z_0 to w_0 and that satisfies $\nu(m_0(z)) = \mu(z)$ for all $z \in \Omega_{z_0,R}$. We define

$$\rho^* = \sup\{\rho\,;\, d^\rho_{\mu,\nu}(z_0,w_0) = 0\};$$

clearly $\rho^* \ge R$. The theorem will be proved if we show that $\rho^* = \infty$. We shall do this by contradiction, i.e. we assume $\rho^* < \infty$, and then derive a contradiction.

So let's assume $\rho^* < \infty$. Consider Ω_{z_0,ρ^*}, the hyperbolic disk around z_0 of radius ρ^*. (See Figure 1 for illustration.) Set $\varepsilon = (R-\varrho)/2$, and consider the points on the hyperbolic circle $C = \partial\Omega_{z_0,\rho^*-\varrho-\varepsilon}$. For every $z_1 \in C$, consider the hyperbolic disk $\Omega_{z_1,\varepsilon/2}$; by Lemma 3.13 there exists a point z_2 in this disk and a corresponding point $w_2 \in \mathcal{D}$ such that $d^R_{\mu,\nu}(z_2,w_2) = 0$, i.e. such that

$$\int_{\Omega_{z_2,R}} |\mu(z) - m'^*\nu(z)|^2\, d\mathrm{vol}_H(z) \ = \ 0$$

for some Möbius transformation m' that maps z_2 to w_2; in particular, we have that

$$\mu(z) = \nu(m'(z)) \ \text{ for all } \ z \in \Omega_{z_2,R} \ . \tag{3.8}$$

The hyperbolic distance from z_2 to $\partial\Omega_{z_0,\rho^*}$ is at least $\varrho + \varepsilon/2$. It follows that the hyperbolic disk $\Omega_{z_2,\varrho+\varepsilon/4}$ is completely contained in Ω_{z_0,ρ^*}; since $\mu(z) = \nu(m_0(z))$ for all $z \in \Omega_{z_0,\rho^*}$, this must therefore hold, in particular, for all $z \in \Omega_{z_2,\varrho+\varepsilon/4}$. Since $\Omega_{z_2,\varrho+\varepsilon/4} \subset \Omega_{z_2,R}$, we also have $\mu(z) = \nu(m'(z))$ for all $z \in \Omega_{z_2,\varrho+\varepsilon/4}$, by (3.8). This implies $\nu(w) = \nu(m_0 \circ (m')^{-1}(w))$ for all $w \in \Omega_{w_2,\varrho+\varepsilon/4}$. Because $\mathcal{N}$ is singly ϱ-$_H$fittable, it follows that $m_0 \circ (m')^{-1}$ must be the identity, or $m_0 = m'$. Combining this with (3.8), we have thus shown that $\mu(z) = \nu(m_0(z))$ for all $z \in \Omega_{z_2,R}$.

Since the distance between z_2 and z_1 is at most $\varepsilon/2$, we also have

$$\Omega_{z_2,R} \supset \Omega_{z_1,R-\varepsilon/2} = \Omega_{z_1,\varrho+3\varepsilon/2} \ .$$

This implies that if we select such a point $z_2(z_1)$ for each $z_1 \in C$, then $\Omega_{z_0,\rho^*-\varrho-\varepsilon} \cup \big(\cup_{z_1\in C}\, \Omega_{z_2(z_1),R}\big)$ covers the open disk $\Omega_{z_0,\rho^*+\varepsilon/2}$. By our earlier argument, $\mu(z) = \nu(m_0(z))$ for all z in each of the $\Omega_{z_2(z_1),R}$; since the same is true on $\Omega_{z_0,\rho^*-\varrho-\varepsilon}$, it follows that $\mu(z) = \nu(m_0(z))$ for all z in $\Omega_{z_0,\rho^*+\varepsilon/2}$. This contradicts the

definition of ρ^* as the supremum of all radii for which this was true; it follows that our initial assumption, that ρ^* is finite, cannot be true, completing the proof. □

For $(\mathcal{D}, \mu)$ to be singly ϱ-$_{\mathrm{H}}$fittable, no two hyperbolic disks $\Omega_{z,R}$, $\Omega_{w,R}$ (where w can equal z) can be isometric via a Möbius transformation m, if $R > \varrho$, except if $m = Id$. However, if z is close (in the Euclidean sense) to the boundary of $\mathcal{D}$, the hyperbolic disk $\Omega_{z,R}$ is very small in the Euclidean sense, and corresponds to a very small piece (near the boundary) of $\mathcal{M}$. This means that single ϱ-$_{\mathrm{H}}$fittability imposes restrictions in increasingly small scales near the boundary of $\mathcal{M}$; from a practical point of view, this is hard to check, and in many applications, the behavior of $\mathcal{M}$ close to its boundary is irrelevant. For this reason, we also formulate the following relaxation of the results above.

Definition 3.15. A surface $\mathcal{M}$ is said to be a singly A-$_{\mathcal{M}}$fittable (where $A > 0$) if there are no patches (i.e. open, path-connected sets) in $\mathcal{M}$ of area larger than A that are isometric, with respect to the metric on $\mathcal{M}$.

If a surface is singly A-$_{\mathcal{M}}$fittable, then it is obviously also A'-$_{\mathcal{M}}$fittable for all $A' \geq A$; the condition of being A-$_{\mathcal{M}}$fittable becomes more restrictive as A decreases. The following theorem states that two singly A-$_{\mathcal{M}}$fittable surfaces at zero $\mathbf{d}^R$-distance from each other must necessarily be isometric, up to some small boundary layer.

Theorem 3.16. *Consider two surfaces $\mathcal{M}$ and $\mathcal{N}$, with corresponding conformal factors μ and ν on $\mathcal{D}$, and suppose $\mathbf{d}^R(\mathcal{M},\mathcal{N}) = 0$ for some $R > 0$. Then the following holds: for arbitrarily large $\rho > 0$, there exist a Möbius transformation $m \in M_D$ and a value $A > 0$ such that if $\mathcal{M}$ and $\mathcal{N}$ are singly A-$_{\mathcal{M}}$fittable then $\mu(m(z)) = \nu(z)$, for all $z \in \Omega_{0,\rho}$.*

Proof. Part of the proof follows the same lines as for Theorem 3.14. We highlight here only the new elements needed for this proof.

First, note that, for arbitrary $r > 0$ and $z_0 \in \mathcal{D}$,

$$\mathrm{vol}_{\mathcal{M}}(\Omega_{z_0,r}) = \int_{\Omega_{z_0,r}} \mu(z) d\mathrm{vol}_H(z) \geq \mathrm{vol}_H(\Omega_{z_0,r}) \left[\min_{z \in \Omega_{z_0,r}} \mu(z) \right] = \mathrm{vol}_H(\Omega_{0,r}) \left[\min_{z \in \Omega_{z_0,r}} \mu(z) \right]. \tag{3.9}$$

This motivates the definition of the sets $\mathcal{O}_{A,r}$,

$$\mathcal{O}_{A,r} = \left\{ z \in \mathcal{D} \;\middle|\; \min_{z' \in \Omega_{z,r}} \mu(z') > \frac{A}{\mathrm{vol}_H(0, \Omega_{0,r})} \right\}; \tag{3.10}$$

$A > 0$ is still arbitrary at this point; its value will be set below.

Now pick $r < R$, and set $\varepsilon = (R - r)/2$. Note that if $z \in \mathcal{O}_{A,r}$, then $\mathrm{vol}_{\mathcal{M}}(\Omega_{z,R}) \geq \mathrm{vol}_{\mathcal{M}}(\Omega_{z,r}) > A$.

Since μ is bounded below by a strictly positive constant on each $\Omega_{0,\rho'}$, we can pick, for arbitrarily large ρ, $A > 0$ such that $\Omega_{0,\rho} \subset \mathcal{O}_{A,r}$; for this it suffices that A exceed a threshold depending on ρ and r. (Since $\mu(z) \to 0$ as z approaches the boundary of $\mathcal{D}$ in Euclidean norm, we expect this threshold to tend towards 0 as $\rho \to \infty$.) We assume that $\Omega_{0,\rho} \subset \mathcal{O}_{A,r}$ in what follows.

Similar to the proof of Theorem 3.14, we invoke Lemma 3.13 to infer the existence of z_0, w_0 such that $z_0 \in \Omega_{0,\varepsilon/2}$ and $d^R_{\mu,\nu}(z_0, w_0) = 0$. We denote

$$\rho^* = \sup\{r' \,;\, d^{r'}_{\mu,\nu}(z_0, w_0) = 0\};$$

as before, there exists a Möbius transformation m such that $\nu(m(z)) = \mu(z)$ for all z in Ω_{z_0,ρ^*}. To complete our proof it therefore suffices to show that $\rho^* \geq \rho + \varepsilon/2$, since $\Omega_{0,\rho} \subset \Omega_{z_0,\rho+\varepsilon/2}$.

Suppose the opposite is true, i.e. $\rho^* < \rho + \varepsilon/2$. By the same arguments as in the proof of Theorem 3.14, there exists, for each $z_1 \in \partial\Omega_{z_0,\rho^*-r-\varepsilon}$, a point $z_2 \in \Omega_{z_1,\varepsilon/2}$ such that $d^R_{\mu,\nu}(z_2, w_2) = 0$ for some w_2. Since the hyperbolic distance between z_2 and 0 is bounded above by $\varepsilon/2 + \rho^* - r - \varepsilon + \varepsilon/2 < \rho - r + \varepsilon/2 < \rho$, $z_2 \in \Omega_{0,\rho} \subset \mathcal{O}_{A,r}$, so that $\mathrm{vol}_{\mathcal{M}}(\Omega_{z_2,R}) > A$. It then follows from the conditions on $\mathcal{M}$ and $\mathcal{N}$ that $\nu(m(z)) =$

$\mu(z)$ for all z in $\Omega_{z_0,\rho^*} \cup \Omega_{z_2,R} \supset \Omega_{z_0,\rho^*} \cup \Omega_{z_1,r+3\varepsilon/2}$. Repeating the argument for all $z_1 \in \partial\Omega_{z_0,\rho^*-r-\varepsilon}$ shows that $\nu(m(z)) = \mu(z)$ can be extended to all $z \in \Omega_{z_0,\rho^*+\varepsilon/2}$, leading to a contradiction that completes the proof. □

4. Discretization and implementation

To transform the theoretical framework constructed in the preceding sections into an algorithm, we need to discretize the relevant continuous objects. Our general plan is to recast the transportation eq. (3.7) as a linear programming problem between discrete measures. This requires two approximation steps:
1) approximating the surface's Uniformization, and
2) discretizing the resulting continuous measures and finding the optimal transport between the discrete measures.

To show how we do this, we first review a few basic notions such as the representation of (approximations to) surfaces by faceted, piecewise flat approximations, called *meshes*, and discrete conformal mappings; the conventions we describe here are the same as adopted in [18].

4.1. Meshes, mid-edge meshes, and discrete conformal mapping.

Triangular (piecewise-linear) meshes are a popular choice for the definition of discrete versions of smooth surfaces. We shall denote a triangular mesh by the triple $M = (V, E, F)$, where $V = \{v_i\}_{i=1}^m \subset \mathbb{R}^3$ is the set of vertices, $E = \{e_{i,j}\}$ the set of edges, and $F = \{f_{i,j,k}\}$ the set of faces (oriented $i \to j \to k$). When dealing with a second surface, we shall denote its mesh by N. We assume our mesh is homeomorphic to a disk.

Next, we introduce "conformal mappings" of a mesh to the unit disk. Natural candidates for discrete conformal mappings are not immediately obvious. Since we are dealing with piecewise linear surfaces, it might seem natural to select a continuous linear maps that is piecewise affine, such that its restriction to each triangle is a similarity transformation. A priori, a similarity map from a triangular face to the disk has 4 degrees of freedom; requiring that the image of each edge remain a shared part of the boundary of the images of the faces abutting the edge, and that the map be continuous when crossing this boundary, imposes 4 constraints for each edge. This quick back of the envelope calculation thus allows $4|F|$ degrees of freedom for such a construction, with $4|E|$ constraints. Since $3|F|/2 \approx |E|$ this problem is over constrained, and a construction along these lines is not possible. A different approach uses the notion of discrete harmonic and discrete conjugate harmonic functions due to Pinkall and Polthier [23, 25] to define a discrete conformal mapping on the mid-edge mesh (to be defined shortly). This relaxes the problem to define a map via a similarity on each triangle that is continuous through only *one* point in each edge, namely the mid point. This procedure was employed in [18]; we will summarize it here; for additional implementation details we refer the interested reader (or programmer) to that paper, which includes a pseudo-code.

The mid-edge mesh $\mathbf{M} = (\mathbf{V}, \mathbf{E}, \mathbf{F})$ of a given mesh $M = (V, E, F)$ is defined as follows. For the vertices $\mathbf{v}_r \in \mathbf{V}$, we pick the mid-points of the edges of the mesh M; we call these the mid-edge points of M. There is thus a $\mathbf{v}_r \in \mathbf{V}$ corresponding to each edge $e_{i,j} \in E$. If $\mathbf{v}_s$ and $\mathbf{v}_r$ are the mid-points of edges in E that share a vertex in M, then there is an edge $\mathbf{e}_{s,r} \in \mathbf{E}$ that connects them. It follows that for each face $f_{i,j,k} \in F$ we can define a corresponding face $\mathbf{f}_{r,s,t} \in \mathbf{F}$, the vertices of which are the mid-edge points of (the edges of) $f_{i,j,k}$; this face has the same orientation as $f_{i,j,k}$. Note that the mid-edge mesh is not a manifold mesh, as illustrated by the mid-edge mesh in Figure 2, shown together with its "parent" mesh: in $\mathbf{M}$ each edge "belongs" to only one face $\mathbf{F}$, as opposed to a manifold mesh, in which most edges (the edges on the boundary are exceptions) function as a hinge between two faces. This "lace" structure makes a mid-edge mesh more flexible: it turns out that it is possible to define a piecewise linear map that makes each face in $\mathbf{F}$ undergo a pure scaling (i.e. all its edges are shrunk or extended by the same factor) and that simultaneously flattens the whole mid-edge mesh. By extending this back to the original mesh, we thus obtain a map from each triangular face to a similar triangle in the plane; these individual similarities can be "knitted together" through the mid-edge points, which continue to coincide (unlike most of the vertices of the original triangles).

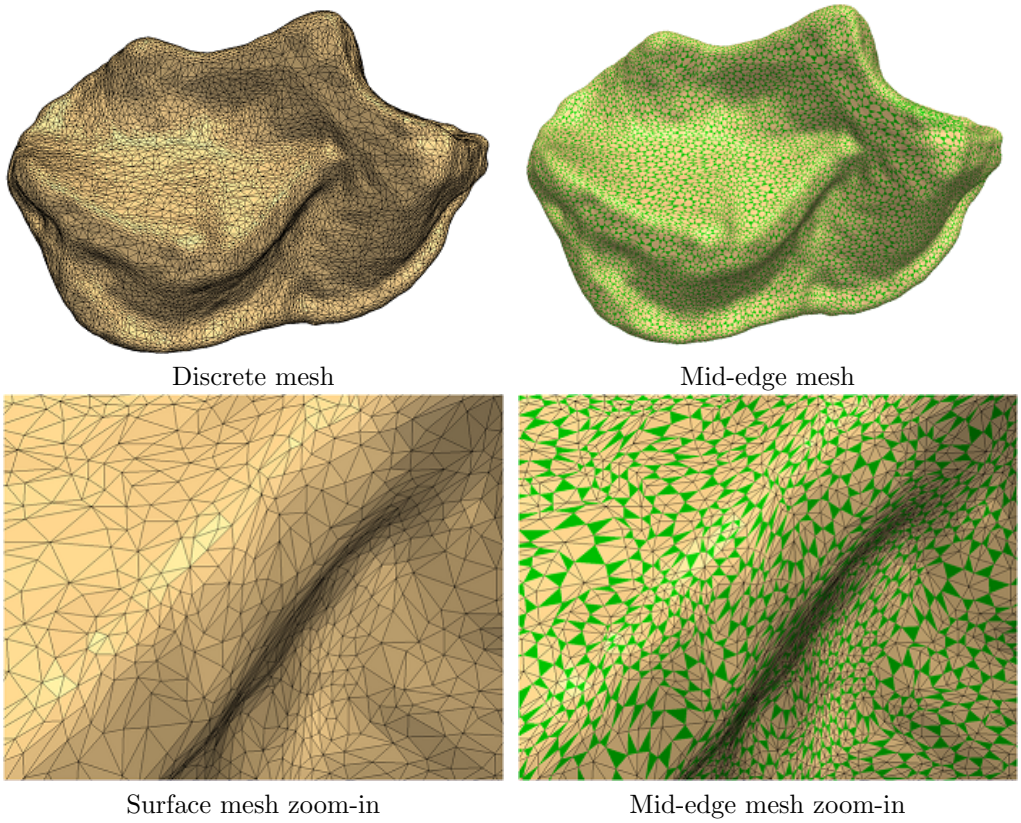


FIGURE 2. A mammalian tooth surface mesh, with the corresponding mid-edge mesh.

To determine the flattening map, we use the framework of discrete harmonic and conjugate harmonic functions, first defined and studied by Pinkall and Polthier [23, 25] in the context of discrete minimal surfaces. This framework was first adapted to the present context in [18]; this adaptation is explained in some detail in Appendix B. The flattening map is well-defined at the mid-edges $\mathbf{v}_s$. As shown in [18] (see also Appendix B) the boundary of the mesh gets mapped onto a region with a straight horizontal slit (see Figure 3, where the boundary points are marked in red). We can assume, without loss of generality, that this slit coincides with the interval $[-2, 2] \subset \mathbb{C}$, since it would suffice to shift and scale the whole figure to make this happen. The holomorphic map $z = w + \frac{1}{w}$ maps the unit disk conformally to $\mathbb{C} \setminus [-2, 2]$, with the boundary of the disk mapped to the slit at $[-2, 2]$; when the inverse of this map is applied to our flattened mid-edge mesh, its image will thus be a mid-edge mesh in the unit disk, with the boundary of the disk corresponding to the boundary of our (disk-like) surface. (See Figure 3.) We shall denote by $\Phi : \mathbf{V} \to \mathbb{C}$ the concatenation of these different conformal and discrete-conformal maps, from the original mid-edge mesh to the corresponding mid-edge mesh in the unit disk.

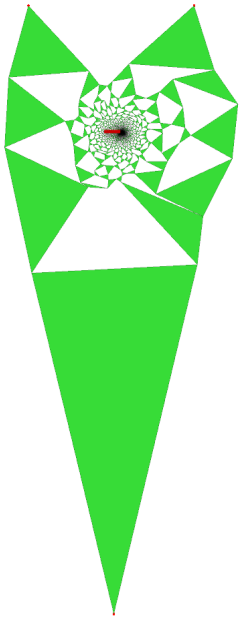
Next, we define the Euclidean discrete conformal factors, defined as the density, w.r.t. the Euclidean metric, of the mid-edge triangles (faces), i.e.

$$\mu^E_{\mathbf{f}_{r,s,t}} = \frac{\mathrm{vol}_{\mathbb{R}^3}(\mathbf{f}_{r,s,t})}{\mathrm{vol}(\Phi(\mathbf{f}_{r,s,t}))}.$$

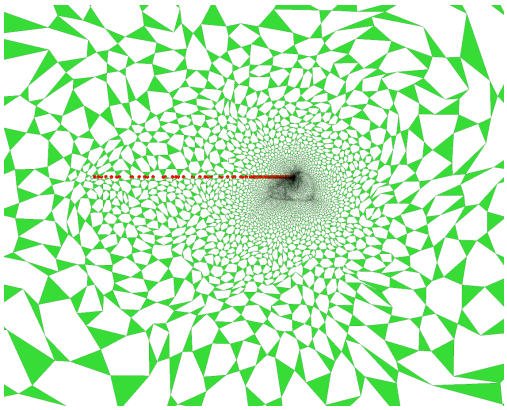
Note that according to this definition, we have

$$\int_{\Phi(\mathbf{f}_{r,s,t})} \mu^E_{\mathbf{f}_{r,s,t}} \, d\mathrm{vol}_E = \frac{\mathrm{vol}_{\mathbb{R}^3}(\mathbf{f}_{r,s,t})}{\mathrm{vol}_E\left(\Phi(\mathbf{f}_{r,s,t})\right)} \mathrm{vol}_E\left(\Phi(\mathbf{f}_{r,s,t})\right) = \mathrm{vol}_{\mathbb{R}^3}(\mathbf{f}_{r,s,t}),$$

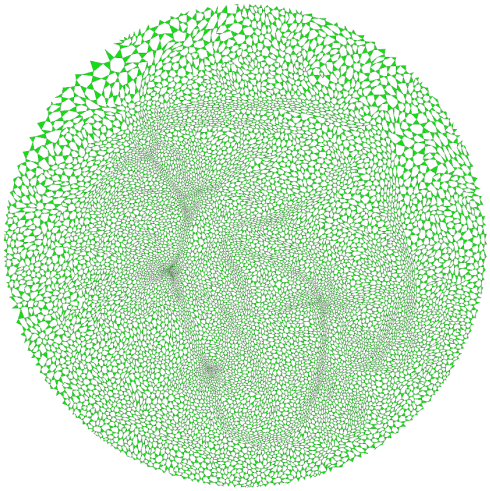
where vol_E denotes the standard Euclidean volume element $dx^1 \wedge dx^2$ in $\mathbb{C}$, and $\mathrm{vol}_{\mathbb{R}^3}(\mathbf{f})$ stands for the area of $\mathbf{f}$ as induced by the standard Euclidean volume element in $\mathbb{R}^3$. The discrete Euclidean conformal factor



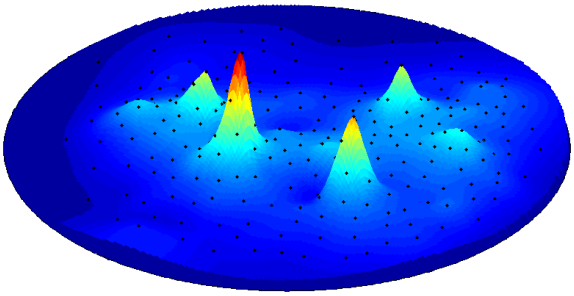
Mid-edge uniformization



Uniformization Zoom-in



After mapping to the disk



Interpolated conformal factor

Figure 3. The discrete conformal transform to the unit disk for the surface of Figure 2, and the interpolation of the corresponding discrete conformal factors (plotted with the JET color map in Matlab). The red points in the top row's images show the boundary points of the disk.

at a mid-edge vertex $\mathbf{v}_r$ is then defined as the average of the conformal factors for the two faces $\mathbf{f}_{r,s,t}$ and $\mathbf{f}_{r,s',t'}$ that touch in $\mathbf{v}_r$, i.e.

$$\mu^E_{\mathbf{v}_r} = \frac{1}{2}\left(\mu^E_{\mathbf{f}_{r,s,t}} + \mu^E_{\mathbf{f}_{r,s',t'}}\right).$$

Figure 3 illustrates the values of the Euclidean conformal factor for the mammalian tooth surface of earlier figures. The discrete hyperbolic conformal factors are defined according to the following equation, consistent

with the convention adopted in section 2,

$$\mu^H_{\mathbf{v}_r} \;=\; \mu^E_{\mathbf{v}_r}\,\left(1-|\Phi(\mathbf{v}_r)|^2\right)^2. \tag{4.1}$$

As before, we shall often drop the superscript: unless otherwise stated, $\mu=\mu^H$, and $\nu=\nu^H$.

The (approximately) conformal mapping of the original mesh to the disk is completed by constructing a smooth interpolant $\Gamma_\mu:\mathcal{D}\to\mathbb{R}$, which interpolates the discrete conformal factor so far defined only at the vertices in $\Phi(\mathbf{V})$; Γ_ν is constructed in the same way. In practice we use Thin-Plate Splines, i.e. functions of the type

$$\Gamma_\mu(z)=p_1(z)\;+\;\sum_i b_i\,\psi(|z-z_i|)\,,$$

where $\psi(r)=r^2\log(r^2)$, $p_1(z)$ is a linear polynomial in x^1,x^2, and $b_i\in\mathbb{C}$; p_1 and the b_i are determined by the data that need to be interpolated. Similarly $\Gamma_\nu(w)\;=\;q_1(w)+\sum_j c_j\,\psi(|w-w_j|)$ for some constants $c_j\in\mathbb{C}$ and a linear polynomial $q_1(w)$ in y^1,y^2. We use as interpolation centers two point sets $Z=\{z_i\}_{i=1}^n$, and $W=\{w_j\}_{j=1}^p$ defined in the next subsection for the discretization of measures. See Figure 3 (bottom-right) the interpolated conformal factor based on the black point set.

We also note that for practical purposes it is sometimes advantageous to use Smoothing Thin-Plate Splines:

$$\Gamma_\mu(z)=\operatorname*{argmin}_\gamma\left\{\lambda\sum_r|\mu_{\mathbf{v}_r}-\gamma(\Phi(\mathbf{v}_r))|^2+(1-\lambda)\int_{\mathcal{D}}\left(\frac{\partial^2\gamma}{\partial(x^1)^2}\right)^2+\left(\frac{\partial^2\gamma}{\partial x^1x^2}\right)^2+\left(\frac{\partial^2\gamma}{\partial(x^2)^2}\right)^2dx^1\wedge dx^2\right\}.$$

when using these, we picked the value 0.99 for the smoothing factor λ.

4.2. **Discretizing continuous measures and their transport.** In this subsection we indicate how to construct discrete approximations $T^R_{\text{discr.},d}(\xi,\zeta)$ for the distance $\mathbf{d}^R(\mathcal{X},\mathcal{Y})=T^R_d(\xi,\zeta)$ between two surfaces $\mathcal{X}$ and $\mathcal{Y}$, each characterized by a corresponding smooth density on the unit disk $\mathcal{D}$ (ξ for $\mathcal{X}$, ζ for $\mathcal{Y}$). (In practice, we will use the smooth functions Γ_μ and Γ_ν for ξ and ζ.) We shall use discrete optimal transport to construct our approximation $T^R_{\text{discr.},d}(\xi,\zeta)$, based on sampling sets for the surfaces, with convergence to the continuous distance as the sampling is refined.

To quantify how fine a sampling set Z is, we use the notion of *fill distance* $\varphi(Z)$:

$$\varphi(Z)\;:=\;\sup\left\{r>0\;\middle|\;z\in\mathcal{M}:B_g(z,r)\cap Z_h=\emptyset\right\}\;,$$

where $B_g(z,r)$ is the geodesic open ball of radius r centered at z. That is, $\varphi(Z)$ is the radius of the largest geodesic ball that can be fitted on the surface $\mathcal{X}$ without including any point of Z. The smaller $\varphi(Z)$, the finer the sampling set.

Given the smooth density ξ (on $\mathcal{D}$), we discretize it by first distributing n points $Z=\{z_i\}_{i=1}^n$ on $\mathcal{X}$ with $\varphi(Z)=h>0$. For $i=1,\dots,n$, we define the sets Ξ_i to be the Voronoi cells corresponding to $z_i\in Z$; this gives a partition of the surface $\mathcal{X}$ into disjoint convex sets, $\mathcal{X}=\cup_{i=1}^n\Xi_i$. We next define the discrete measure ξ_Z as a superposition of point measures localized in the points of Z, with weights given by the areas of Ξ_i, i.e. $\xi_Z\;=\;\sum_{i=1}^n\xi_i\delta_{z_i}$, with $\xi_i:=\xi(\Xi_i)=\int_{\Xi_i}d\text{vol}_{\mathcal{X}}$. Similarly we denote by $W=\{w_j\}_{j=1}^p$, Υ_j, and $\zeta_j:=\zeta(\Upsilon_j)$ the corresponding quantities for surface $\mathcal{Y}$. We shall always assume that the surfaces $\mathcal{X}$ and $\mathcal{Y}$ have the same area, which, for convenience, we can take to be 1. It then follows that the discrete measures ξ_Z and ζ_W have equal total mass (regardless of whether $n=p$ or not). The approximation algorithm will compute optimal transport for the discrete measures ξ_Z and ζ_W; the corresponding discrete approximation to the distance between ξ and ζ is then given by $T^R_d(\xi_Z,\zeta_W)$.

Convergence of the discrete approximations $T^R_d(\xi_Z,\zeta_W)$ to $T^R_d(\xi,\zeta)=\mathbf{d}^R(\mathcal{X},\mathcal{Y})$ as $\varphi(Z)$, $\varphi(W)\to0$ then follows from the results proved in [17]. Corollary 3.3 in [17] requires that the distance function $d^R_{\xi,\zeta}(\cdot,\cdot)$ used to define $T^R_d(\xi,\zeta)$ be uniformly continuous in its two arguments. We can establish this in our present case by invoking the continuity properties of $d^R_{\xi,\zeta}$ proved in Theorem 3.7, extended by the following lemma, proved in Appendix A.

Lemma 4.1. *Let $\{(z_k, w_k)\}_{k\geq 1} \subset \mathcal{D}\times\mathcal{D}$ be a sequence that converges, in the Euclidean norm, to some point in $(z', w') \in \overline{\mathcal{D}} \times \overline{\mathcal{D}} \setminus \mathcal{D} \times \mathcal{D}$, that is $|z_k - z'| + |w_k - w'| \to 0$, as $k \to \infty$. Then, $\lim_{k\to\infty} d^R_{\xi,\zeta}(z_k, w_k)$ exists and depends only on the limit point (z', w').*

We shall denote this continuous extension of $d^R_{\mu,\nu}(\cdot,\cdot)$ to $\overline{\mathcal{D}} \times \overline{\mathcal{D}}$ by the same symbol $d^R_{\mu,\nu}$.

Since $\overline{\mathcal{D}} \times \overline{D}$ is compact, (this extension of) $d^R_{\xi,\zeta}(\cdot,\cdot)$ is uniformly continuous: for all $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that, for all $z, z' \in \mathcal{X}$, $w, w' \in \mathcal{Y}$,

$$d_{\mathcal{X}}(z, z') < \delta(\varepsilon)\ , d_{\mathcal{Y}}(w, w') < \delta(\varepsilon)\ \Rightarrow\ \left|d^R_{\xi,\zeta}(z, w) - d^R_{\xi,\zeta}(z', w')\right| < \varepsilon,$$

where $d_{\mathcal{X}}(\cdot,\cdot)$ is the geodesic distance on $\mathcal{X}$, and $d_{\mathcal{Y}}(\cdot,\cdot)$ is the geodesic distance on $\mathcal{Y}$.

The results in [17] then imply that $\xi_Z \to \xi$ in the *weak* sense, as $\varphi(Z) \to 0$, i.e. that for all bounded continuous functions $f : \overline{\mathcal{D}} \to \mathbb{R}$, the convergence $\int_{\mathcal{D}} f\ d\xi_Z \to \int_{\mathcal{D}} f\ d\xi$ holds [22]. Similarly and $\zeta_W \to \zeta$ in the weak sense as $\varphi(W) \to 0$. Furthermore, [17] also proves that for $\max(\varphi(Z), \varphi(W)) < \frac{\delta(\varepsilon)}{2}$

$$\left|T^R_d(\xi_Z, \zeta_W) - T^R_d(\xi, \zeta)\right| < \varepsilon.$$

More generally, it is shown that

$$\left|T^R_d(\xi_Z, \zeta_W) - T^R_d(\xi, \zeta)\right| < \omega_{d^R_{\xi,\zeta}}\left(\max(\varphi(Z), \varphi(W))\right), \tag{4.2}$$

where $\omega_{d^R_{\xi,\zeta}}$ is the modulus of continuity of $d^R_{\xi,\zeta}$, that is

$$\omega_{d^R_{\xi,\zeta}}(t) = \sup_{d_{\mathcal{X}}(z,z')+d_{\mathcal{Y}}(w,w')<t} \left|d^R_{\xi,\zeta}(z, w) - d^R_{\xi,\zeta}(z', w')\right|.$$

We shall see below that it will be particularly useful to choose the centers in $Z = \{z_i\}_{i=1}^n$, $W = \{w_j\}_{j=1}^p$ such that the corresponding Voronoi cells are (approximately) of equal area, i.e. $n = N = p$ and $\xi_i = \xi(\Xi_i) \approx \frac{1}{N}$, $\zeta_j = \zeta(\Upsilon_j) \approx \frac{1}{N}$, where we have used that the total area of each surface is normalized to 1. An effective way to calculate such sample sets Z and W is to start from an initial random seed (which will not be included in the set), and take the geodesic point furthest from the seed as the initial point of the sample set. One then keeps repeating this procedure, selecting at each iteration the point that lies at the furthest geodesic distance from the set of points already selected. This algorithm is known as the Farthest Point Algorithm (FPS) [7]. An example of the output of this algorithm, using geodesic distances on a disk-type surface, is shown in Figure 4. Further discussion of practical aspects of Voronoi sampling of a surface can be found in [4].

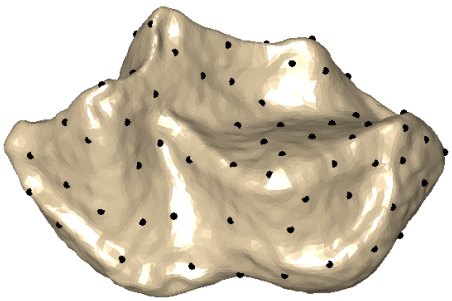


FIGURE 4. Sampling of the surface of Figure 2 obtained by the Farthest Point Algorithm.

4.3. **Approximating the local distance function $d^R_{\mu,\nu}$.** We are now ready to construct our discrete version of the optimal volume transportation for surfaces (3.7). The previous subsection describes how to derive the discrete measures μ_Z, ν_W from the approximate conformal densities Γ_μ, Γ_ν and the sampling sets Z and W. For simplicity, we will, with some abuse of notation, identify the approximations Γ_μ, Γ_ν with μ, ν. The approximation error made here is typically much smaller than the errors made in further

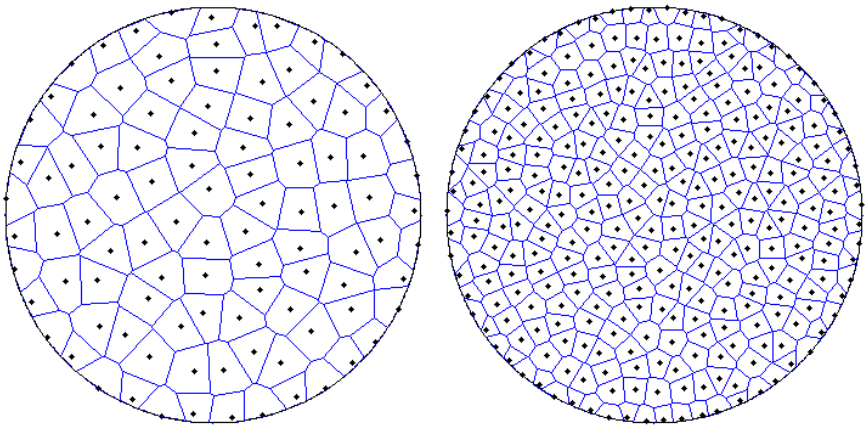


FIGURE 5. The integration centers and their corresponding Voronoi cells used for calculating the integration weights for the discrete quadrature. Left: 100 centers; Right: 300.

steps (see below) and we shall neglect it. The final component is approximating $d^R_{\mu,\nu}(z_i, w_j)$ for all pairs $(z_i, w_j) \in Z \times W$. Applying (3.4) to the points z_i, w_j we have:

$$d^R_{\mu,\nu}(z_i, w_j) = \min_{m(z_i)=w_j} \int_{\Omega_{z_i,R}} \Big| \mu(z) - \nu(m(z)) \Big| \, d\mathrm{vol}_H. \tag{4.3}$$

To obtain $d^R_{\mu,\nu}(z_i, w_j)$ we will thus need to compute integrals over hyperbolic disks of radius R, which is done via a separate approximation procedure, set up once and for all in a preprocessing step at the start of the algorithm.

By using a Möbius transformation $\widetilde{m}$ such that $\widetilde{m}(0) = z_0$, and the identity

$$\int_{\Omega_{0,R}} \Big| \mu(\widetilde{m}(u)) - \nu(m \circ \widetilde{m}(u)) \Big| \, d\mathrm{vol}_H(u) = \int_{\Omega_{z_0,R}} \Big| \mu(z)) - \nu(m(z)) \Big| \, d\mathrm{vol}_H(z) \ ,$$

we can reduce the integrals over the hyperbolic disks $\Omega_{z_i,R}$ to integrals over a hyperbolic disk centered around zero.

In order to (approximately) compute integrals over $\Omega_0 = \Omega_{0,R} = \{z| \, |z| \leq r_R\}$, we first pick a positive integer K and distribute centers p_k, $k = 1, ..., K$ in Ω_0. We then decompose Ω_0 into Voronoi cells Δ_k corresponding to the p_k, obtaining $\Omega_0 = \cup_{k=1}^K \Delta_k$; see Figure 5 (note that these Voronoi cells are completely independent of those used in 4.2.)

To approximate the integral of a continuous function f over Ω_0 we then use

$$\int_{\Omega_0} f(z) \, d\mathrm{vol}_H(z) \approx \sum_k \left[\int_{\Delta_k} d\mathrm{vol}_H(z) \right] f(p_k) = \sum_k \alpha_k f(p_k)$$

where $\alpha_k = \int_{\Delta_k} d\mathrm{vol}_H(z)$.

We thus have the following approximation:

$$\begin{aligned} d^R_{\mu,\nu}(z_i, w_j) &= \min_{m(z_i)=w_j} \int_{\Omega_{z_i,R}} \Big| \mu(z) - \nu(m(z)) \Big| \, d\mathrm{vol}_H(z) \\ &= \min_{m(z_i)=w_j} \int_{\Omega_{0,R}} \Big| \mu(\widetilde{m}_i(z)) - \nu(m(\widetilde{m}_i(z))) \Big| \, d\mathrm{vol}_H(z) \\ &\approx \min_{m(z_i)=w_j} \sum_k \alpha_k \Big| \mu(\widetilde{m}_i(p_k)) - \nu(m(\widetilde{m}_i(p_k))) \Big| \ , \end{aligned} \tag{4.4}$$

where the Möbius transformations $\widetilde{m}_i$, mapping 0 to z_i, are selected as soon as the z_i themselves have been picked, and remain the same throughout the remainder of the algorithm.

It can be shown that picking a set of centers $\{p_k\}$ with fill-distance $h > 0$ leads to an $O(h)$ approximation; in Appendix A we prove:

Theorem 4.2. *For continuously differentiable μ, ν,*

$$\left| d^R_{\mu,\nu}(z_i, w_j) - \min_{m(z_i)=w_j} \sum_k \alpha_k \, |\, \mu(\widetilde{m}_i(p_k)) - \nu(m(\widetilde{m}_i(p_k)))\,| \right| \leq C\,\varphi\left(\{p_k\}\right) \ ,$$

where the constant C depends only on μ, ν, R.

Let us denote this approximation by

$$\widehat{d}^R_{\mu,\nu}(z_i, w_j) = \min_{m(z_i)=w_j} \sum_k \alpha_k \, |\, \mu(\widetilde{m}_i(p_k)) - \nu(m(\widetilde{m}_i(p_k)))\,| \, .$$

Since the above theorem guarantees that the approximation error $\left|\widehat{d}^R_{\mu,\nu}(z_i, w_j) - d^R_{\mu,\nu}(z_i, w_j)\right|$ can be uniformly bounded independently of z_i, w_j, it can be shown that

$$\left| T^R_d(\mu_Z, \nu_W) - T^R_{\widehat{d}}(\mu_Z, \nu_W) \right| \leq C\varphi\left(\{p_k\}\right),$$

where again C is dependent only upon μ, ν, R. Combining this with eq.(4.2) we get that

$$\left| T^R_d(\mu, \nu) - T^R_{\widehat{d}}(\mu_Z, \nu_W) \right| \leq \omega_{d^R_{\mu,\nu}}\left(\max\left(\varphi(Z), \varphi(W)\right)\right) + C\varphi\left(\{p_k\}\right). \tag{4.5}$$

In practice, for calculating $\widehat{d}^R_{\mu,\nu}$, the minimization over M_{D,z_i,w_j}, the set of all Möbius transformations that map z_i to w_k, is discretized as well: instead of minimizing over all $m_{z_i,w_j,\sigma}$ (see subsection 3.1), we minimize over only the Möbius transformations $\left(m_{z_i,w_j,2\pi\ell/L}\right)_{\ell=0,1,..,L-1}$. Taking this into account as well, we have thus

$$d^R_{\mu,\nu}(z_i, w_j) \approx \min_{\ell=1,\ldots L} \sum_k \alpha_k \left| \mu(\widetilde{m}_i(p_k)) - \nu(m_{z_i,w_j,2\pi\ell/L}(\widetilde{m}_i(p_k))) \right| \, ; \tag{4.6}$$

the error made in approximation (4.6) is therefore proportional to $L^{-1} + C\varphi\left(\{p_k\}\right)$.

To summarize, our approximation $T^R_{\widehat{d}}(\mu_Z, \nu_W)$ to the uniformly continuous $T^R_d(\mu, \nu)$ is based on two approximations: on the one hand, we compute the transportation cost between the discrete measures μ_Z, ν_W, approximating μ, ν; on the other hand, this transportation cost involves a local distance $\widehat{d}^R_{\mu,\nu}$ which is itself an approximation. The transportation between the discrete measures will be computed by solving a linear programming optimization, as explained in detail in the next subsection. The final approximation error (4.5) depends on two factors: 1) the fill distances $\varphi(Z), \varphi(W)$ of the sample sets Z, W, and 2) the approximation of the local distance function $d^R_{\mu,\nu}(z_i, w_j)$ between the sample points. Combining the discretization of the Möbius search with (4.6), the total approximation error is thus proportional to $\omega_{d^R_{\Gamma_\mu,\Gamma_\nu}}\left(\varphi\left(\{p_k\}\right)\right) + L^{-1} + \varphi\left(\{p_k\}\right)$.

Recall that we are in fact using Γ_μ, Γ_ν in the role of of μ, ν (see above), which entails an additional approximation error. This error relates to the accuracy with which discrete meshes approximate smooth manifolds, as well as the method used to approximate uniformization. We come back to this question in Appendix B. As far as we are aware, a full convergence result for (any) discrete uniformization is still unknown; in any case, we expect this error to be negligible (and approximately of the order of the largest edge in the full mesh) compared to the others.

4.4. **Optimization via linear programming.** The discrete formulation of eq. (3.7) is commonly formulated as follows:

$$\sum_{i,j} d_{ij}\pi_{ij} \to \min \tag{4.7}$$

$$\begin{cases} \sum_i \pi_{ij} = \nu_j \\ \sum_j \pi_{ij} = \mu_i \\ \pi_{ij} \geq 0 \end{cases}, \tag{4.8}$$

where $\mu_i = \mu(\Xi_i)$ and $\nu_j = \nu(\Upsilon_j)$, and $d_{ij} = d^R_{\mu,\nu}(z_i, w_j)$.

In practice, surfaces are often only partially isometric (with a large overlapping part), or the sampled points may not have a good one-to-one and onto correspondence (i.e. there are points both in Z and in W that do not correspond well to any point in the other set). In these cases it is desirable to allow the algorithm to consider transportation plans π with marginals *smaller or equal* to μ and ν. Intuitively this means that we allow that only some fraction of the mass is transported and that the remainder can be "thrown away". This leads to the following formulation:

$$\sum_{i,j} d_{ij}\pi_{ij} \to \min \tag{4.9}$$

$$\begin{cases} \sum_i \pi_{ij} \leq \nu_j \\ \sum_j \pi_{ij} \leq \mu_i \\ \sum_{i,j} \pi_{ij} = Q \\ \pi_{ij} \geq 0 \end{cases} \tag{4.10}$$

where $0 < Q \leq 1$ is a parameter set by the user that indicates how much mass *must* be transported, in total.

The corresponding transportation distance is defined by

$$T_d(\nu, \nu) = \sum_{ij} d_{ij}\pi_{ij}, \tag{4.11}$$

where π_{ij} are the entries in the matrix π for the optimal (discrete) transportation plan.

Since these equations and constraints are all linear, we have the following theorem:

Theorem 4.3. *The equations* (4.7)-(4.8) *and* (4.9)- (4.10) *admit a global minimizer that can be computed in polynomial time, using standard linear-programming techniques.*

When correspondences between surfaces are sought, i.e. when one imagines one surface as being transformed into the other, one is interested in restricting π to the class of permutation matrices instead of allowing all bistochastic matrices. (This means that each entry π_{ij} is either 0 or 1.) In this case the number of centers z_i must equal that of w_j, i.e. $n = N = p$, and it is best to pick the centers so that $\mu_i = \frac{1}{N} = \nu_j$, for all i, j. It turns out that this is sufficient to *guarantee* (without restricting the choice of π in any way) that the minimizing π is a permutation:

Theorem 4.4. *If $n = N = p$ and $\mu_i = \frac{1}{N} = \nu_j$, then*

1. *There exists a global minimizer of* (4.7) *that is a permutation matrix.*
2. *If furthermore $Q = \frac{M}{N}$, where $M < N$ is an integer, then there exists a global minimizer of* (4.9) *π such that $\pi_{ij} \in \{0, 1\}$ for each i, j.*

Remark 4.5. In the second case, where $\pi_{ij} \in \{0, 1\}$ for each i, j and $\sum_{i,j=1}^{N} \pi_{ij} = M$, π can still be viewed as a permutation of M objects, "filled up with zeros". That is, if the zero rows and columns of π (which must exist, by the pigeon hole principle) are removed, then the remaining $M \times M$ matrix is a permutation.

Proof. We first note that in both cases, we can simply renormalize each μ_i and ν_j by N, leading to the rescaled systems

$$\left\{\begin{array}{c}\sum_i \pi_{ij} = 1\\ \sum_j \pi_{ij} = 1\\ \pi_{ij} \geq 0\end{array}\right. \qquad \left\{\begin{array}{c}\sum_i \pi_{ij} \leq 1\\ \sum_j \pi_{ij} \leq 1\\ \sum_{i,j} \pi_{ij} = M\\ \pi_{ij} \geq 0\end{array}\right. \tag{4.12}$$

To prove the first part, we note that the left system in (4.12) defines a convex polytope in the vector space of matrices that is exactly the Birkhoff polytope of bistochastic matrices. By the Birkhoff-Von Neumann Theorem [15] every bistochastic matrix is a convex combination of the permutation matrices, i.e. each π satisfying the left system in (4.12) must be of the form $\sum_k c_k \tau^k$, where the τ^k are the $N!$ permutation matrices for N objects, and $\sum_k c_k = 1$, with $c_k \geq 0$. The minimizing π in this polytope for the linear functional (4.7) must thus be of this form as well. It follows that at least one τ^k must also minimize (4.7), since otherwise we would obtain the contradiction

$$\sum_{ij} d_{ij}\pi_{ij} = \sum_k c_k \Big(\sum_{ij} d_{ij}\tau^k_{ij} \Big) \geq \min_k \Big\{ \sum_{ij} d_{ij}\tau^k_{ij} \Big\} > \sum_{i,j} d_{ij}\,\pi_{ij}\ . \tag{4.13}$$

The second part can be proved along similar steps: the right system in (4.12) defines a convex polytope in the vector space of matrices; it follows that every matrix that satisfies the system of constraints is a convex combination of the extremal points of this polytope. It suffices to prove that these extreme points are exactly those matrices that satisfy the constraints and have entries that are either 0 or 1 (this is the analog of the Birkhoff-von Neumann theorem for this case; we prove this generalization in a lemma in Appendix C); the same argument as above then shows that there must be at least one extremal point where the linear functional (4.7) attains its minimum. □

This means that, when we seek correspondences between two surfaces, there is no need to *impose* the (very nonlinear) constraint on π that it be a permutation matrix; one can simply use a linear program to solve either , with Theorem 4.4 guaranteeing that the minimizer for the "relaxed" problem (4.7)-(4.8) or (4.9)-(4.10) is of the desired type if $n = N = p$ and $\mu_i = \frac{1}{N} = \nu_j$.

4.5. **Consistency.** In our schemes to compute the surface transportation distance, for example by solving (4.9), we have so far not included any constraints on the regularity of the resulting optimal transportation plan π^*. When computing the distance between a surface and a reasonable deformation of the same surface, one does indeed find, in practice, that the minimizing π^* is fairly smooth, because neighboring points have similar neighborhoods. There is no guarantee, however, that this has to happen. Moreover, we will be interested in comparing surfaces that are far from (almost) isometric, given by noisy datasets. Under such circumstances, the minimizing π^* may well "jump around". In this subsection we propose a regularization procedure to avoid such behavior.

Computing how two surfaces best correspond makes use of the values of the "distances in similarity" $d^R_{\mu,\nu}(z_i, w_j)$ between pairs of points that "start" on one surface and "end" on the other; computing these values relies on finding a minimizing Möbius transformation for the functional (3.4). We can keep track of these minimizing Möbius transformations m_{ij} for the pairs of points (z_i, w_j) proposed for optimal correspondence by the optimal transport algorithm described above. Correspondence pairs (i, j) that truly participate in some close-to-isometry map will typically have Möbius transformations m_{ij} that are very similar. This suggests a method of filtering out possibly mismatched pairs, by retaining only the set of correspondences (i, j) that cluster together within the Möbius group.

There exist many ways to find clusters. In our applications, we gauge how far each Möbius transformation m_{ij} is from the others by computing a type of ℓ_1 variance:

$$E_V(i, j) = \sum_{(k,\ell)} \|m_{ij} - m_{k\ell}\|\,, \tag{4.14}$$

where the norm is the Frobenius norm (also called the Hilbert-Schmidt norm) of the 2×2 complex matrices representing the Möbius transformations, after normalizing them to have determinant one. We then use $E_V(i,j)$ as a consistency measure of the corresponding pair (i,j).

5. Examples and comments

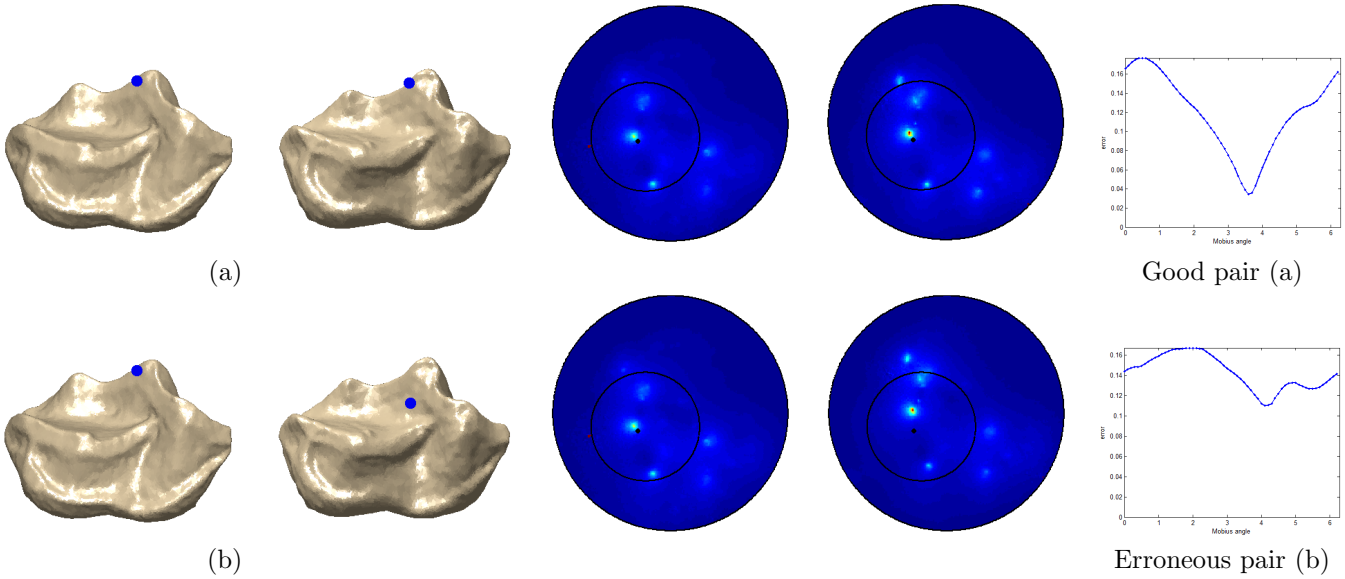


FIGURE 6. Calculation of the local distance $d^R_{\mu,\nu}(\cdot,\cdot)$ between pairs of points on two different surfaces (each row shows a different pair of points; the two surfaces are the same in the top and bottom rows). The first row shows a "good" pair of points together with the alignment of the conformal densities $\mu, m^*\nu$ based on the best Möbius transformation m minimizing $\int_{\mathcal{D}} \|\mu - m^*\nu\| \, d\mathrm{vol}_{\mathcal{M}}$. The plot of this latter integral as a function of m (parameterized by $\sigma \in [0, 2\pi)$, see (2.3)) is shown in the right-most column. The second row shows a "bad" correspondence which indeed leads to a higher local distance $d^R_{\mu,\nu}$.

In this section we present a few experimental results using our new surface comparison operator. These concern an application to biology; in a case study of the use of our approach to the characterization of mammals by the surfaces of their molars, we compare high resolution scans of the masticating surfaces of molars of several lemurs, which are small primates living in Madagascar. Traditionally, biologists specializing in this area carefully determine landmarks on the tooth surfaces, and measure characteristic distances and angles involving these landmarks. A first stage of comparing different tooth surfaces is to identify correspondences between landmarks. Figure 6 illustrates how $d^R_{\mu,\nu}(z,w)$ can be used to find corresponding pairs of points on two surfaces by showing both a "good" and a "bad" corresponding pair. The left two columns of the figure show the pair of points in each case; the two middle columns show the best fit after applying the minimizing Möbius on the corresponding disk representations; the rightmost column plots $\int_{\Omega_{z_0,R}} |\,\mu(z) - (m^*_{z_0,w_0,\sigma}\nu)(z)\,| \, d\mathrm{vol}_H(z)$, the value of the "error", as a function of parameter σ, parametrizing the Möbius transformations that map a give point z_0 to another given point w_0 (see Lemma 3.5). The "best" corresponding point w_0 for a given z_0 is the one that produces the lowest minimal value for the error, i.e. the lowest $d^R_{\mu,\nu}(z_0,w_0)$.

Figure 7 show the top 120 most consistent corresponding pairs (in groups of 20) for two molars belonging to lemurs of different species. Corresponding pairs are indicated by highlighted points of the same color. These correspondences have surprised the biologists from whom we obtained the data sets; their experimental measuring work, which incorporates finely balanced judgment calls, had defied earlier automatizing attempts.

Once the differences and similarities between molars from different animals have been quantified, they can be used (as part of an approach) to classify the different individuals. Figure 8 illustrates a preliminary result from [13] that illustrates the possibility of such classifications based on the distance operator between surfaces introduced in this paper. The figure illustrates the pairwise distance matrix for eight molars, coming from individuals in four different species (indicated by color). The clustering was based on only the distances between the molar surfaces; it clearly agrees with the clustering by species, as communicated to us by the biologists from whom we obtained the data sets.

One final comment regarding the computational complexity of our method. There are two main parts: the preparation of the distance matrix d_{ij} and the linear programming optimization. For the linear programming part we used a Matlab interior point implementation with N^2 unknowns, where N is the number of points spread on the surfaces. In our experiments, the optimization typically terminated after $15-20$ iterations for $N = 150-200$ points, which took about 2-3 seconds. The computation of the similarity distance d_{ij} took longer, and was the bottleneck in our experiments. If we spread N points on each surface, and use them all (which was usually not necessary) to interpolate the conformal factors Γ_μ, Γ_ν, if we use P points in the integration rule, and take L points in the Möbius discretization (see Section 4 for details) then each approximation of $d^R_{\mu,\nu}(z_i, w_j)$ by (4.6) requires $O(L \cdot P \cdot N)$ calculations, as each evaluation of Γ_μ, Γ_ν takes $O(N)$ and we need $L \cdot P$ of those. Since we have $O(N^2)$ distances to compute, the computation complexity for calculating the similarity distance matrix d_{ij} is $O(L \cdot P \cdot N^3)$. In practice this step was the most time consuming and took around two hours for $N = 300$. However, we have not used any code optimization and we believe these times can be reduced significantly.

6. Acknowledgments

The authors would like to thank Cédric Villani and Thomas Funkhouser for valuable discussions, and Jesus Puente for helping with the implementation. We are grateful to Jukka Jernvall, Stephen King, and Doug Boyer for providing us with the tooth data sets, and for many interesting comments. ID gratefully acknowledges (partial) support for this work by NSF grant DMS-0914892, and by an AFOSR Complex Networks grant; YL thanks the Rothschild foundation for postdoctoral fellowship support.

References

1. R. Kimmel A. M. Bronstein, M. M. Bronstein, *Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching*, Proc. National Academy of Sciences (PNAS) **103** (2006), no. 5, 1168–1172.
2. Mikael Fortelius Jukka Jernvall Alistair R. Evans, Gregory P. Wilson, *High-level similarity of dentitions in carnivorans and rodents*, Nature **445** (2007), 78–81.
3. Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, third ed., Texts in applied mathematics, vol. 15, 2008.
4. Alexander Bronstein, Michael Bronstein, and Ron Kimmel, *Calculus of nonrigid surfaces for geometry and texture manipulation*, IEEE Transactions on Visualization and Computer Graphics **13** (2007), no. 5, 902–913.
5. E. Cela, *The quadratic assignment problem: Theory and algorithms (combinatorial optimization)*, Springer, 1998.
6. Gerhard Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, vol. 1357, Springer Berlin / Heidelberg, 1988.
7. Y. Eldar, M. Lindenbaum, M. Porat, and Y. Zeevi, *The farthest point strategy for progressive image sampling*, 1997.
8. Bruce Fischl, Martin I. Sereno, Roger B. H. Tootell, and Anders M. Dale, *High-resolution intersubject averaging and a coordinate system for the cortical surface*, Hum. Brain Mapp **8** (1999), 272–284.
9. Mikhail Gromov, M. Katz, P. Pansu, and S. Semmes, *Metric structures for riemannian and non-riemannian spaces*, Birkhäuser Boston, December 2006.
10. Xianfeng Gu and Shing-Tung Yau, *Global conformal surface parameterization*, SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing (Aire-la-Ville, Switzerland, Switzerland), Eurographics Association, 2003, pp. 127–137.
11. Irwin Kra Hershel M. Farkas, *Riemann surfaces*, Springer, 1992.
12. Hildebrandt, Klaus, Polthier, Konrad, Wardetzky, and Max, *On the convergence of metric and geometric properties of polyhedral surfaces*, Geometriae Dedicata **123** (2006), no. 1, 89–112.
13. Jesus Puente Ingrid Daubechies, Yaron Lipman, *Metric analysis of the disc-durface manifold*, In preparation (2010).
14. L. Kantorovich, *On the translocation of masses*, C.R. (Dokl.) Acad. Sci. URSS (N.S.) **37** (1942), 199–201.

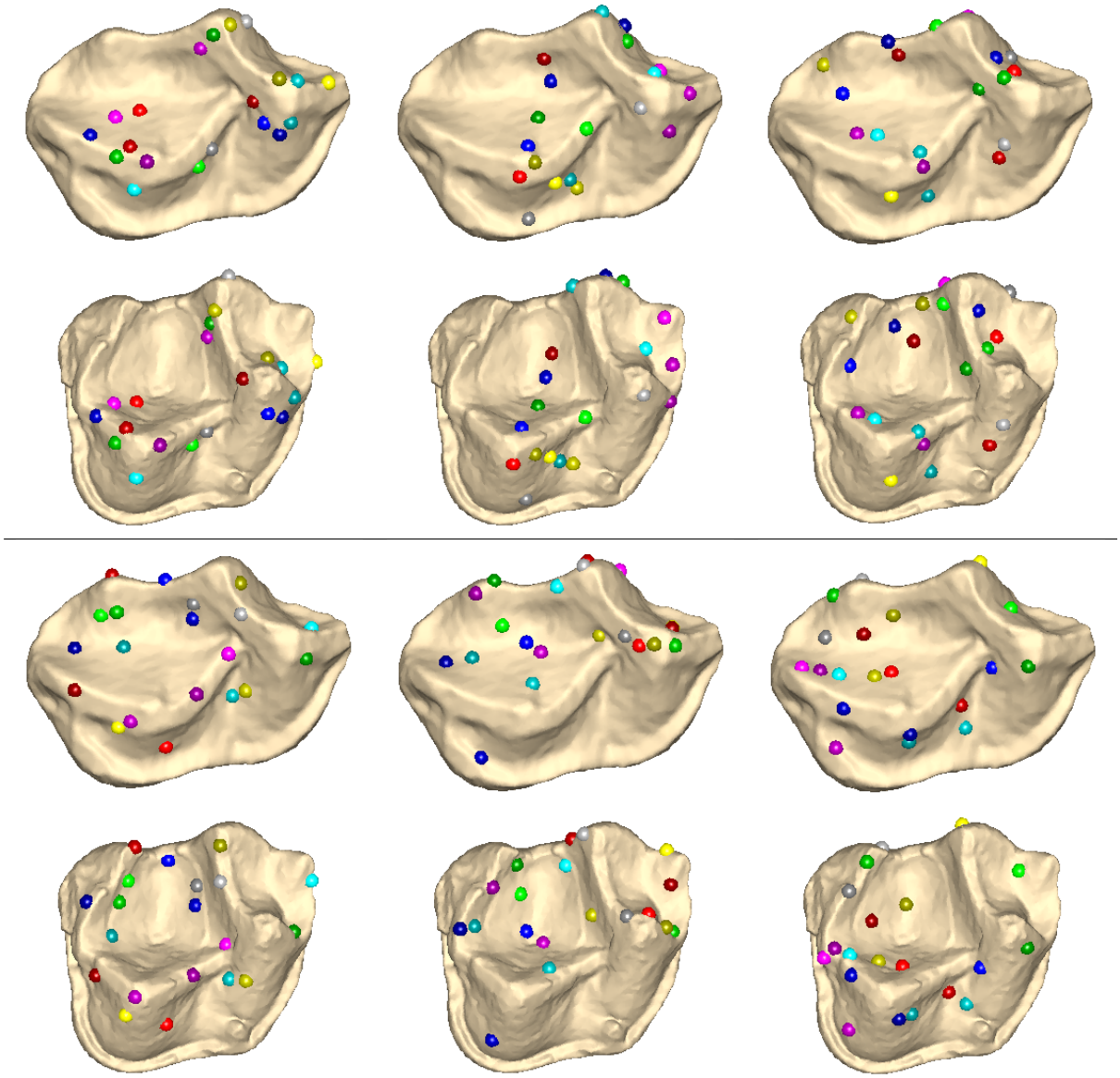


Figure 7. The top 120 most consistent corresponding pairs between two molar teeth models.

15. M.D. Plummer L. Lovász, *Matching theory*, North-Holland, 1986.
16. Parsons LM Liotti M Freitas CS Rainey L Kochunov PV Nickerson D Mikiten SA Fox PT Lancaster JL, Woldorff MG, *Automated talairach atlas labels for functional brain mapping*, Human Brain Mapping **10** (2000), 120–131.
17. Yaron Lipman, *Approximation of optimal transport cost*, Technical report (2009).
18. Yaron Lipman and Thomas Funkhouser, *Mobius voting for surface correspondence*, ACM Transactions on Graphics (Proc. SIGGRAPH) **28** (2009), no. 3.
19. Facundo Memoli, *On the use of gromov-hausdorff distances for shape comparison*, Symposium on Point Based Graphics (2007).
20. Facundo Mémoli and Guillermo Sapiro, *A theoretical and computational framework for isometry invariant recognition of point cloud data*, Found. Comput. Math. **5** (2005), no. 3, 313–347.
21. G. Monge, *Mmoire sur la thorie des dblais et de remblais*, Histoire de lAcadmie Royale des Sciences de Paris, avec les Mmoires de Mathmatique et de Physique pour la mme anne (1781), 666–704.
22. Billingsley Patrick, *Convergence of probability measures*, John Wiley & Sons, 1968.

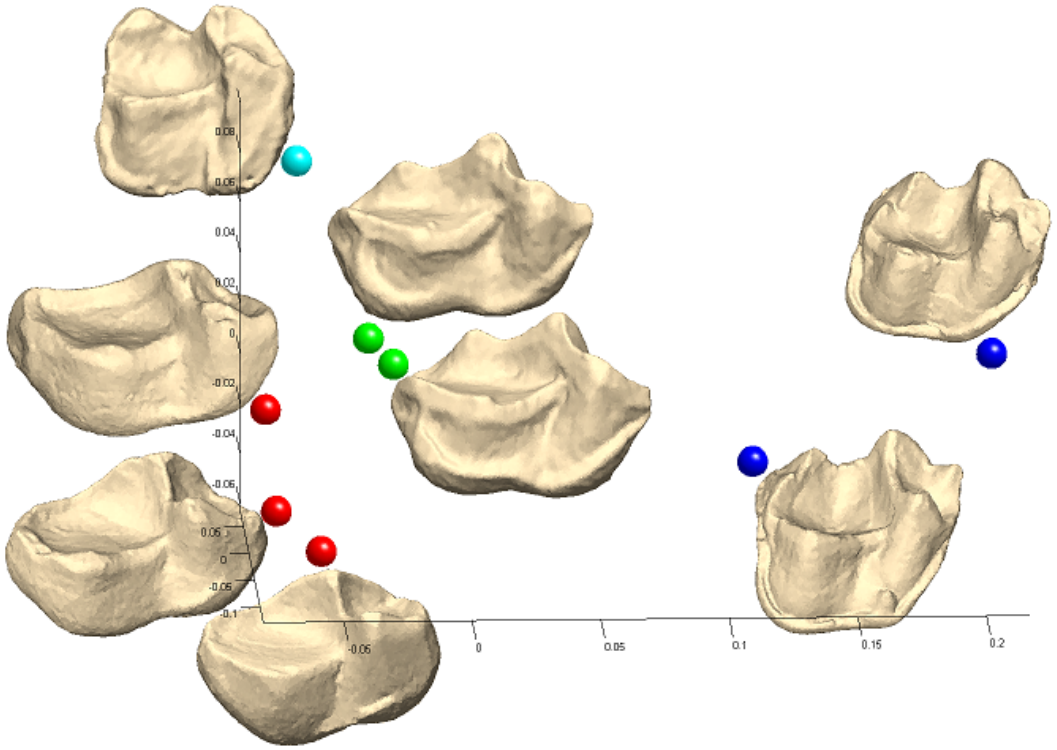


Figure 8. Embedding of the distance graph of eight teeth models using multi-dimensional scaling. Different colors represent different lemur species. The graph suggests that the geometry of the teeth might suffice to classify species.

23. Ulrich Pinkall and Konrad Polthier, *Computing discrete minimal surfaces and their conjugates*, Experimental Mathematics **2** (1993), 15–36.
24. Konrad Polthier, *Conjugate harmonic maps and minimal surfaces*, Preprint No. 446, TU-Berlin, SFB 288 (2000).
25. ———, *Computational aspects of discrete minimal surfaces*, Global Theory of Minimal Surfaces, Proc. of the Clay Mathematics Institute 2001 Summer School, David Hoffman (Ed.), CMI/AMS (2005).
26. Y. Rubner, C. Tomasi, and L. J. Guibas, *The earth mover's distance as a metric for image retrieval*, International Journal of Computer Vision **40** (2000), no. 2, 99–121.
27. Conroy B Bryan RE Ramadge PJ Haxby JV Sabuncu MR, Singer BD, *Function-based intersubject alignment of human cortical anatomy*, Cereb Cortex. (2009).
28. Alexander Schrijver, *A course in combinatorial optimization, course note*, 2008.
29. George Springer, *Introduction to riemann surfaces*, AMS Chelsea Publishing, 1981.
30. Cedric Villani, *Topics in optimal transportation (graduate studies in mathematics, vol. 58)*, American Mathematical Society, March 2003.
31. W. Zeng, X. Yin, Y. Zeng, Y. Lai, X. Gu, and D. Samaras, *3d face matching and registration based on hyperbolic ricci flow*, CVPR Workshop on 3D Face Processing (2008), 1–8.
32. W. Zeng, Y. Zeng, Y. Wang, X. Yin, X. Gu, and D. Samaras, *3d non-rigid surface matching and registration based on holomorphic differentials*, The 10th European Conference on Computer Vision (ECCV) (2008).

Appendix A.

This Appendix contains some technical proofs of Lemmas and Theorems stated in section 3, and 4. We start with proving the list of properties of the distance function $d^R_{\mu,\nu}(z,w)$ given in Theorem 3.3:

Theorem 3.3

The distance function $d^R_{\mu,\nu}(z,w)$ satisfies the following properties

(1)	$d^R_{m_1^*\mu,m_2^*\nu}(m_1^{-1}(z_0),m_2^{-1}(w_0)) = d^R_{\mu,\nu}(z_0,w_0)$	Invariance under (well-defined) Möbius changes of coordinates
(2)	$d^R_{\mu,\nu}(z_0,w_0) = d^R_{\nu,\mu}(w_0,z_0)$	Symmetry
(3)	$d^R_{\mu,\nu}(z_0,w_0) \geq 0$	Non-negativity
(4)	$d^R_{\mu,\nu}(z_0,w_0) = 0 \implies \Omega_{z_0,R}$ in $(\mathcal{D},\mu)$ and $\Omega_{w_0,R}$ in $(\mathcal{D},\nu)$ are isometric	
(5)	$d^R_{m^*\nu,\nu}(m^{-1}(z_0),z_0) = 0$	Reflexivity
(6)	$d^R_{\mu_1,\mu_3}(z_1,z_3) \leq d^R_{\mu_1,\mu_2}(z_1,z_2) + d^R_{\mu_2,\mu_3}(z_2,z_3)$	Triangle inequality

Proof. For (1), denote $m_1^{-1}(z_0) = z_1$, and $m_2^{-1}(w_0) = w_1$. Then

$$\begin{aligned} d^R_{m_1^*\mu,m_2^*\nu}(z_1,w_1) &= \inf_{m(z_1)=w_1} \int_{\Omega_{z_1,R}} |m_1^*\mu(z) - m^*m_2^*\nu(z)| d\mathrm{vol}_H(z) \\ &= \inf_{m(z_1)=w_1} \int_{\Omega_{z_1,R}} |\mu(m_1(z)) - \nu(m_2(m(z)))| d\mathrm{vol}_H(z). \end{aligned}$$

Next set $\widetilde{m} = m_2 \circ m \circ m_1^{-1}$. Note that $\widetilde{m}(z_0) = w_0$. Plugging $m_2(m(z)) = \widetilde{m}(m_1(z))$ into the integral and carrying out the change of variables $m_1(z) = z'$, we obtain

$$\inf_{m(z_1)=w_1} \int_{\Omega_{z_1,R}} |\,\mu(z') - \nu(\widetilde{m}(z'))\,| d\mathrm{vol}_H(z') = \inf_{\widetilde{m}(z_0)=w_0} \int_{\Omega_{z_0,R}} |\,\mu(z') - \nu(\widetilde{m}(z'))\,|\, d\mathrm{vol}_H(z').$$

For (2), we use Lemma 3.2 and equations (2.5), (2.6) to write

$$\begin{aligned} d^R_{\mu,\nu}(z_0,w_0) &= \inf_{m(z_0)=w_0} \int_{\Omega_{z_0,R}} |\mu(z) - m^*\nu(z)| d\mathrm{vol}_H(z) \\ &= \inf_{m(z_0)=w_0} \int_{\Omega_{w_0,R}} |(m^{-1})^*\mu(w) - \nu(w)| d\mathrm{vol}_H(w) = d^R_{\nu,\mu}(w_0,z_0). \end{aligned}$$

(3) and (4) are immediate from the definition of $d^R_{\mu,\nu}$.

(5) follows from the observation that the minimizing m (in the definition (3.4) of $d^R_{\mu,\nu}$) is m_1 itself, for which the integrand, and thus the whole integral vanishes identically.

For (6), let m_1 be a Möbius transformation such that $m_1(z_1) = z_2$, and m_2 such that $m_2(z_2) = z_3$. Setting $m = m_2 \circ m_1$, we have

$$\begin{aligned} d^R_{\mu_1,\mu_3}(z_1,z_3) &\leq \int_{\Omega_{z_1,R}} |\,\mu_1(z) - m^*\mu_3(z)\,|\, d\mathrm{vol}_H(z) \\ &\leq \int_{\Omega_{z_1,R}} |\,\mu_1(z) - m_1^*\mu_2(z)\,|\, d\mathrm{vol}_H(z) + \int_{\Omega_{z_1,R}} |\,m_1^*\mu_2(z) - m^*\mu_3(z)\,|\, d\mathrm{vol}_H(z)\,. \end{aligned} \tag{A.1}$$

The second term in (A.1) can be rewritten as (using Lemma 3.2, the change of coordinates $m_1(z_1) = z_2$ and the observation $m^* = m_1^* m_2^*$)

$$\begin{aligned}\int_{\Omega_{z_1,R}} |\, m_1^*\mu_2(z) - m^*\mu_3(z)\,|\, d\mathrm{vol}_H(z) &= \int_{\Omega_{z_2,R}} |\, m_{1*}m_1^*\mu_2(w) - m_{1*}m_1^*m_2^*\mu_3(w)\,|\, d\mathrm{vol}_H(w)\\ &= \int_{\Omega_{z_2,R}} |\,\mu_2(w) - m_2^*\mu_3(w)\,|\, d\mathrm{vol}_H(w).\end{aligned}$$

We have thus

$$d^R_{\mu_1,\mu_3}(z_1,z_3) \le \int_{\Omega_{z_1,R}} |\,\mu_1(z) - m_1^*\mu_2(z)\,|\, d\mathrm{vol}_H(z) + \int_{\Omega_{z_2,R}} |\,\mu_2(w) - m_2^*\mu_3(w)\,|\, d\mathrm{vol}_H(w)\ ,$$

and this for any m_1, $m_2 \in M_D$ such that $m_1(z_1) = z_2$ and $m_2(z_2) = z_3$. Minimizing over m_1 and m_2 then leads to the desired result.

□

Next we prove the continuity properties of the function $\Phi(z_0, w_0, \sigma) = \int_{\Omega(z_0,R)} |\,\mu(z) - \nu(m_{z_0,w_0,\sigma}(z))\,|\, d\mathrm{vol}_H(z)$, stated in Lemma 3.6, which were used to prove continuity of $d^R_{\mu,\nu}$ itself (in Theorem 3.7).

Lemma 3.6

- *For each fixed (z_0, w_0) the function $\Phi(z_0, w_0, \cdot)$ is continuous on S_1.*
- *For each fixed $\sigma \in S_1$, $\Phi(\cdot,\cdot,\sigma)$ is continuous on $\mathcal{D}\times\mathcal{D}$. Moreover, the family $\Big(\Phi(\cdot,\cdot,\sigma)\Big)_{\sigma\in S_1}$ is equicontinuous.*

Proof. We start with the continuity in σ. We have

$$|\,\Phi(z_0,w_0,\sigma) - \Phi(z_0,w_0,\sigma')\,| \le \int_{\Omega(z_0,R)} |\nu(m_{z_0,w_0,\sigma}(z)) - \nu(m_{z_0,w_0,\sigma'}(z))\,|\, d\mathrm{vol}_H(z)\ .$$

Because ν is continuous on $\mathcal{D}$, its restriction to the compact set $\overline{\Omega(w_0,R)}$ (the closure of $\Omega(w_0,R)$) is bounded. Since the hyperbolic volume of $\Omega(z_0,R)$ is finite, the integrand is dominated, uniformly in σ', by an integrable function. Since $m_{z_0,w_0,\sigma}(z)$ is obviously continuous in σ, we can use the dominated convergence theorem to conclude.

Since S^1 is compact, this continuity implies that the infimum in the definition of $d^R_{\mu,\nu}$ can be replaced by a minimum:

$$d^R_{\mu,\nu}(z_0,w_0) = \min_{m(z_0)=w_0} \int_{\Omega(z_0,R)} |\,\mu(z) - \nu(m(z))\,|\, d\mathrm{vol}_H(z)\ .$$

Next we prove continuity in z_0 and w_0 (with estimates that are uniform in σ).

Consider two pairs of points, (z_0,w_0) and $(z_0',w_0') \in \mathcal{D}\times\mathcal{D}$. Then

$$\begin{aligned}&|\,\Phi(z_0,w_0,\sigma) - \Phi(z_0',w_0',\sigma)\,|\\ &= \left| \int_{\Omega(z_0,R)} |\,\mu(z) - \nu(m_{z_0,w_0,\sigma}(z))\,|\, d\mathrm{vol}_H(z) - \int_{\Omega(z_0',R)} |\,\mu(u) - \nu(m_{z_0',w_0'\sigma}(u)\,|\, d\mathrm{vol}_H(u) \right|\\ &\le \left| \int_{\Omega(z_0,R)} |\,\mu(z) - \nu(m_{z_0,w_0,\sigma}(z))\,|\, d\mathrm{vol}_H(z) - \int_{\Omega(z_0,R)} |\,\mu(m_{z_0,z_0',1}z) - \nu(m_{z_0',w_0',\sigma}\circ m_{z_0,z_0',1}(z))\,|\, d\mathrm{vol}_H(u) \right|\\ &\le \int_{\Omega(z_0,R)} \big(\,|\,\mu(z) - \mu(m_{z_0,z_0',1}(z))\,| + |\,\nu(m_{z_0,w_0,\sigma}(z)) - \nu(m_{z_0',w_0',\sigma}(m_{z_0,z_0',1}(z)))\,|\,\big)\ d\mathrm{vol}_H(z)\ .\end{aligned}$$

On the other hand, note that for any $\gamma > 0$, μ and ν are continuous on the closures of $\Omega(z_0, R+\gamma)$ and $\Omega(w_0, R+\gamma)$, respectively; since these closed hyperbolic disks are compact, μ and ν are bounded on these sets. Pick now $\rho > 0$ such that $|z_0' - z_0| < \rho$, $|w_0' - w_0| < \rho$ imply that $\Omega(z_0', R) \subset \Omega(z_0, R+\gamma)$ as well as

$\Omega(w_0', R) \subset \Omega(w_0, R+\gamma)$. It follows that, if $|z_0' - z_0| < \rho$ and $|w_0' - w_0| < \rho$, then $|\,\mu(z) - \mu(m_{z_0,z_0',1}(z)\,|$ and $|\,\nu(m_{z_0 0,w_0,\sigma}(z)) - \nu(m_{z_0',w_0',\sigma}(m_{z_0,z_0',1}(z)))\,|$ are bounded uniformly for $z \in \Omega(z_0, R)$. Since it is clear from the explicit expressions (3.6) that $m_{z_0,z_0',1}(z) \to z$ and $m_{z_0',w_0',\sigma}(m_{z_0,z_0',1}(z)) \to m_{z_0,w_0,\sigma}(z)$ as $z_0' \to z_0$ and $w_0' \to w_0$, we can thus invoke the dominated convergence theorem again to prove continuity of $\Phi(\cdot,\cdot,\sigma)$.

To prove the equicontinuity, we first note that ν is uniformly continuous on $\Omega(w_0, R) \cup \Omega(w_0', R)$, since ν is continuous on the compact set $\overline{\Omega(w_0, R+\gamma)}$, which contains $\Omega(w_0, R) \cup \Omega(w_0', R)$ for all w_0' that satisfy $|w_0' - w_0| \le \rho$. This means that, given any $\varepsilon > 0$, we can find $\delta > 0$ such that $|\nu(w) - \nu(w')| \le \varepsilon$ holds for all w, w' that satisfy w, $w' \in \Omega(w_0, R) \cup \Omega(w_0', R)$ and $|w - w'| \le \delta$. This implies the desired equicontinuity if we can show that $|m_{z_0,w_0,\sigma}(z) - m_{z_0',w_0',\sigma}(m_{z_0,z_0',1}(z))|$ can be made smaller than δ, uniformly in $\sigma \in S_1$, by making $|z_0' - z_0| + |w_0' - w_0|$ sufficiently small.
We first estimate $|m_{z_0,w_0,\sigma}(z) - m_{z_0,w_0',\sigma}(z)|$. With the notations of (3.6), we have

$$\begin{aligned} a(z_0, w_0, \sigma) - a(z_0, w_0', \sigma) &= \frac{(z_0 - w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma}) - (z_0 - w_0'\overline{\sigma})(1 - \overline{z_0}w_0\overline{\sigma})}{(1 - \overline{z_0}w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma})} \\ &= \frac{(w_0 - w_0')\overline{\sigma}(|z_0|^2 - 1)}{(1 - \overline{z_0}w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma})}\ , \end{aligned}$$

so that

$$|a(z_0, w_0, \sigma) - a(z_0, w_0', \sigma)| \le \frac{|w_0 - w_0'|}{(1 - |z_0|\,|w_0|)[1 - |z_0|(|w_0| + \xi)]} \le \frac{\xi}{(1 - |z_0|\,|w_0|)[1 - |z_0|(|w_0| + \xi)]}$$

when $|w_0 - w_0'| < \xi$. It thus suffices to choose ξ so that $\xi < \zeta(1 - |z_0|\,|w_0|)[1 - |z_0|(|w_0| + \xi)]$ to ensure that $|a(z_0, w_0, \sigma) - a(z_0, w_0', \sigma)| < \zeta$. For the phase factor τ in (3.6) we obtain

$$\begin{aligned} \tau(z_0, w_0, \sigma) - \tau(z_0, w_0', \sigma) &= \sigma\, \frac{(1 - \overline{z_0}w_0'\overline{\sigma})(1 - z_0\overline{w_0}\sigma) - (1 - \overline{z_0}w_0\overline{\sigma})(1 - z_0\overline{w_0'}\sigma)}{(1 - \overline{z_0}w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma})} \\ &= \sigma\, \frac{(w_0 - w_0')\overline{z_0}\overline{\sigma} - (\overline{w_0} - \overline{w_0'})z_0\sigma + |z_0|^2(\overline{w_0}w_0' - \overline{w_0'}w_0)}{(1 - \overline{z_0}w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma})} \\ &= \sigma\, \frac{(w_0 - w_0')\overline{z_0}\overline{\sigma} - z_0(\overline{w_0} - \overline{w_0'})\sigma + |z_0|^2[\overline{w_0}(w_0' - w_0) + w_0(\overline{w_0} - \overline{w_0'})]}{(1 - \overline{z_0}w_0\overline{\sigma})(1 - \overline{z_0}w_0'\overline{\sigma})}\ ; \end{aligned}$$

when $|w_0 - w_0'| < \xi$, this implies

$$|\tau(z_0, w_0, \sigma) - \tau(z_0, w_0', \sigma)| \le \frac{|z_0|\,|w_0|\,[2 + |z_0|(2|w_0| + \xi)]}{(1 - |z_0|\,|w_0|)[1 - |z_0|(|w_0| + \xi)]}\xi\ ,$$

which can clearly be made smaller than any $\zeta > 0$ by choosing ξ sufficiently small. All this implies that (use (3.6))

$$\begin{aligned} |m_{z_0,w_0,\sigma}(z) - m_{z_0,w_0',\sigma}(z)| &\le |\tau(z_0, w_0, \sigma) - \tau(z_0, w_0', \sigma)|\frac{1 + |z|}{1 - |z|}\ +\ |a(z_0, w_0, \sigma) - a(z_0, w_0', \sigma)|\frac{(1 + |z|)^2}{(1 - |z|)^2} \\ &\le \zeta\, \frac{2(1 + |z|)}{(1 - |z|)^2}, \end{aligned}$$

which will be smaller than $\delta/2$, uniformly in σ, if $\zeta < \delta(1 - |z|^2)/8$; this bound on ζ in turn determines the bound to be imposed on the ξ used above. Hence $|m_{z_0,w_0,\sigma}(z) - m_{z_0,w_0',\sigma}(z)| < \delta/2$ can be guaranteed, uniformly in σ, by choosing $|w_0 - w_0'| < \xi$ for sufficiently small ξ.
One can estimate likewise

$$|m_{z_0,w_0',\sigma}(z) - m_{z_0',w_0',\sigma}(m_{z_0',z_0,1}(z))|\ ,$$

and show that this too can be made smaller than $\delta/2$, uniformly in σ, by imposing sufficiently tight bounds on $|z_0' - z_0|$ and $|w_0' - w_0|$. Combining all these estimates then leads to the desired equicontinuity, as indicated earlier. □

To prove Lemma 4.1, we shall use the following lemma:

Lemma A.1. *Consider $u_k = e^{\mathbf{i}\psi} + \varepsilon_k$, where $|\varepsilon_k| \to 0$ as $k \to \infty$. Then there exists, for every $\varepsilon > 0$, a $K \in \mathbb{N}$ such that for all $k > K$; and all $\widehat{m} \in M_{D,0,u_k}$,*

$$\inf_{w \in \Omega_{0,R}} |\widehat{m}(w)| > 1 - \varepsilon.$$

The set $M_{D,0,u_k}$ used in this lemma is given by Definition 3.4.

Proof. From Lemma 3.5 we can write $\widehat{m}$ as

$$\widehat{m}(w) = e^{\mathbf{i}\theta} \frac{w + u_k e^{-\mathbf{i}\theta}}{1 + \overline{u_k} e^{\mathbf{i}\theta} w},$$

for some $\theta \in [0, 2\pi)$. Substituting $u_k = e^{\mathbf{i}\psi} + \varepsilon_k$ in this equation we get

$$\widehat{m}(w) = e^{\mathbf{i}\theta} \frac{w + (e^{\mathbf{i}\psi} + \varepsilon_k) e^{-\mathbf{i}\theta}}{1 + \overline{(e^{\mathbf{i}\psi} + \varepsilon_k)} e^{\mathbf{i}\theta} w} = e^{\mathbf{i}\psi} \frac{1 + w e^{\mathbf{i}(\theta - \psi)} + \varepsilon_k \varepsilon^{-\mathbf{i}\psi}}{1 + w e^{\mathbf{i}(\theta-\psi)} + \overline{\varepsilon_k} e^{\mathbf{i}\theta} w}.$$

Writing the shorthand s for $s = 1 + w e^{\mathbf{i}(\theta - \psi)}$, we have thus

$$\begin{aligned} \left| \widehat{m}(w) - e^{\mathbf{i}\psi} \right| &= \left| e^{\mathbf{i}\psi} \frac{s + \varepsilon_k \varepsilon^{-\mathbf{i}\psi}}{s + \overline{\varepsilon_k} e^{\mathbf{i}\theta} w} - e^{\mathbf{i}\psi} \right| \leq \left| e^{\mathbf{i}\psi} \frac{\varepsilon_k e^{-\mathbf{i}\psi} - \overline{\varepsilon}_k e^{\mathbf{i}\theta} w}{s + \overline{\varepsilon}_k e^{\mathbf{i}\theta} w} \right| \\ &\leq \frac{\left| \varepsilon_k e^{-\mathbf{i}\psi} - \overline{\varepsilon}_k e^{\mathbf{i}\theta} w \right|}{\left| s + \overline{\varepsilon}_k e^{\mathbf{i}\theta} w \right|} \leq \frac{|\varepsilon_k| \, (1 + |w|)}{|s| - |\varepsilon_k||w|} \end{aligned}$$

Now for all $w \in \Omega_{0,R}$, $|w| < r_R = \tanh^{-1}(R)$. This implies $|s| \geq 1 - |w| \geq 1 - r_R$, and $1 + |w| \leq 1 + r_R$, so that

$$\left| \widehat{m}(w) - e^{\mathbf{i}\psi} \right| \leq |\varepsilon_k| \frac{1 + r_R}{1 - r_R - |\varepsilon_k| r_R} = |\varepsilon_k| \frac{1 + r_R}{1 - r_R \, (1 + |\varepsilon_k|)}.$$

Since $|\varepsilon_k| \to 0$ the lemma follows. □

We are now ready for

Lemma 4.1 *Let $\{(z_k, w_k)\}_{k \geq 1} \subset \mathcal{D} \times \mathcal{D}$ be a sequence that converges, in the Euclidean norm, to some point $(z', w') \in \overline{\mathcal{D}} \times \overline{\mathcal{D}} \setminus \mathcal{D} \times \mathcal{D}$, that is $|z_k - z'| + |w_k - w'| \to 0$, as $k \to \infty$. Then, $\lim_{k \to \infty} d^R_{\xi,\zeta}(z_k, w_k)$ exists and depends only on the limit point (z', w').*

Proof. Since $(z', w') \in \overline{\mathcal{D}} \times \overline{\mathcal{D}} \setminus \mathcal{D} \times \mathcal{D}$ either $z' \in \overline{\mathcal{D}} \setminus \mathcal{D}$ or $w' \in \overline{\mathcal{D}} \setminus \mathcal{D}$. Let us assume that $z' \in \overline{\mathcal{D}} \setminus \mathcal{D}$ (the case $w' \in \overline{\mathcal{D}} \setminus \mathcal{D}$ is similar). Denote by m_k an arbitrary Möbius transformation in $M_{D,0,w_k}$. By symmetry of the distance and using a change of variables we then obtain

$$\begin{aligned} d^R_{\xi,\zeta}(z_k, w_k) &= d^R_{\zeta,\xi}(w_k, z_k) \\ &= \min_{m(w_k) = z_k} \int_{\Omega_{w_k,R}} \Big| \zeta(w) - \xi(m(w)) \Big| d\mathrm{vol}_H(w) \\ &= \min_{m(w_k) = z_k} \int_{\Omega_{0,R}} \Big| \zeta(m_k(w)) - \xi(m(m_k(w))) \Big| d\mathrm{vol}_H(w). \end{aligned}$$

Now, recall that $\xi(z) = \xi^H(z) = \widetilde{\xi}(z)(1 - |z|^2)^2$, where $\widetilde{\xi}(z)$ is a bounded function, $\sup_{z \in \mathcal{D}} |\widetilde{\xi}(z)| \leq C_{\widetilde{\xi}}$. From Lemma A.1 we know that for every $\varepsilon > 0$ and for $k > K$ sufficiently large, $|m(m_k((w))| > 1 - \varepsilon$ for all $w \in \Omega_{0,R}$, and all m such that $m(w_k) = z_k$. This means that for these $k > K$ we have

$$\begin{aligned} |\xi(m(m_k(w)))| &= \left| \widetilde{\xi}(m(m_k(w))) \right| (1 - |m(m_k(w))|^2)^2 \\ &\leq C_{\widetilde{\xi}} (1 - (1 - \varepsilon)^2)^2 \leq C_{\widetilde{\xi}} \varepsilon^2 (2 - \varepsilon)^2, \end{aligned}$$

for all $w \in \Omega_{0,R}$. Therefore,

$$\begin{aligned}
&\left| d^R_{\xi,\zeta}(z_k, w_k) - \int_{\Omega_{0,R}} \left|\zeta(m_k(w))\right| d\mathrm{vol}_H(w) \right| \\
&\leq \left| \min_{m(w_k)=z_k} \int_{\Omega_{0,R}} \left|\zeta(m_k(w)) - \xi(m(m_k(w)))\right| d\mathrm{vol}_H(w) - \int_{\Omega_{0,R}} \left|\zeta(m_k(w))\right| d\mathrm{vol}_H(w) \right| \\
&\leq \left| \min_{m(w_k)=z_k} \int_{\Omega_{0,R}} \left\{ \left|\zeta(m_k(w)) - \xi(m(m_k(w)))\right| - \left|\zeta(m_k(w))\right| \right\} d\mathrm{vol}_H(w) \right| \\
&\leq \min_{m(w_k)=z_k} \int_{\Omega_{0,R}} \left|\xi(m(m_k(w)))\right| d\mathrm{vol}_H(w) \to 0, \text{ as } k \to \infty.
\end{aligned}$$

Therefore $d^R_{\xi,\zeta}(z_k, w_k)$ converges, as $k \to \infty$, if and only if $\int_{\Omega_{0,R}} |\zeta(m_k(w))|\, d\mathrm{vol}_H(w)$ converges, and to the same limit, for any $m_k \in M_{D,0,w_k}$. We can take, for instance, $m_k(w) = \frac{w+w_k}{1+\overline{w_k}w}$ which gives

$$\int_{\Omega_{0,R}} |\zeta(m_k(w))|\, d\mathrm{vol}_H(w) = \int_{\Omega_{0,R}} \left| \zeta\left(\frac{w+w_k}{1+\overline{w_k}w}\right) \right| d\mathrm{vol}_H(w).$$

For $w \in \Omega_{0,R}$, $|1+\overline{w_k}w| > 1 - r_R$. It follows that this expression has a limit as $k \to \infty$, and

$$\lim_{k\to\infty} \int_{\Omega_{0,R}} |\zeta(m_k(w))|\, d\mathrm{vol}_H(w) = \int_{\Omega_{0,R}} \left| \zeta\left(\frac{w+w'}{1+\overline{w'}w}\right) \right| d\mathrm{vol}_H(w),$$

which clearly depends on w', not on the sequence $\{w_k\}$. □

Next, we prove Theorem 4.2. We start with a simple lemma showing that all Möbius transformations restricted to $\Omega_{0,R}$, $R < \infty$, are Lipschitz with a universal constant, for which we provide an upper bound.

Lemma A.2. *A Möbius transformation $m \in M_D$ restricted to $\Omega_{0,R}$, $R < \infty$ is Lipschitz continuous with Lipschitz constant $C_m \leq \frac{1-|a|^2}{(1-r_R|a|)^2}$.*

Proof. Denote $m(z) = e^{\mathbf{i}\theta}\frac{z-a}{1-z\overline{a}}$. Then, for $z, w \in \Omega_{0,R}$ we have

$$\begin{aligned}
\left|m(z) - m(w)\right| &\leq \left| e^{\mathbf{i}\theta}\frac{z-a}{1-z\overline{a}} - e^{\mathbf{i}\theta}\frac{w-a}{1-w\overline{a}} \right| \leq \left| \frac{(z-a)(1-w\overline{a}) - (w-a)(1-z\overline{a})}{(1-z\overline{a})(1-w\overline{a})} \right| \\
&\leq \left| \frac{(z-w)(1-|a|^2)}{(1-z\overline{a})(1-w\overline{a})} \right| \leq |z-w| \frac{1-|a|^2}{(1-r_R\,|a|)^2}.
\end{aligned}$$

□

Next we prove:

Theorem 4.2 *For continuously differentiable μ, ν,*

$$\left| d^R_{\mu,\nu}(z_i, w_j) - \min_{m(z_i)=w_j} \sum_k \alpha_k \,|\, \mu(\widetilde{m}_i(p_k)) - \nu(m(\widetilde{m}_i(p_k)))\,| \right| \leq C\,\varphi\left(\{p_k\}\right) \ ,$$

where the constant C depends only on μ, ν, R.

Proof. First, denote $f(z) = \left|\mu(\widetilde{m}_i(z)) - \nu(m(\widetilde{m}_i(z)))\right|$. Then,

$$\begin{aligned}
\left| \int_{\Omega_0} f(z) d\mathrm{vol}_H(z) - \min_{m(z_i)=w_j} \sum_k \alpha_k f(p_k) \right| &\leq \textstyle\sum_k \int_{\Omega_0} \left| f(z) - f(p_k) \right| d\mathrm{vol}_H(z) \\
&\leq \omega_f^{\Omega_0}\left(\varphi\left(\{p_k\}\right)\right) \textstyle\int_{\Omega_0} d\mathrm{vol}_H,
\end{aligned} \tag{A.2}$$

where the modulus of continuity $\omega_f^{\Omega_0}(h) = \sup_{|z-w|<h; z,w\in\Omega_0} |f(z)-f(w)|$ is used. Note that

$$\omega_f^{\Omega_0} \le \omega_{\mu\circ\widetilde{m}_i}^{\Omega_0} + \omega_{\nu\circ m\circ\widetilde{m}_i}^{\Omega_0}. \tag{A.3}$$

Since μ, ν have continuous derivatives on compact domain, they are Lipschitz continuous. Denote their Lipschitz constants by C_μ, C_ν, respectively. From Lemma A.2 we see that, for $z, w \in \Omega_0$,

$$\left|\mu(\widetilde{m}_i(z)) - \mu(\widetilde{m}_i(w))\right| \le C_\mu \left|\widetilde{m}_i(z) - \widetilde{m}_i(w)\right| \le C_\mu \frac{1-|a|^2}{(1-r_R|a|)^2}\,|z-w| \le C_\mu \frac{1}{(1-r_R)^2}\,|z-w|\,,$$

which is independent of $\widetilde{m}_i$. Similarly,

$$\left|\nu(m(\widetilde{m}_i(z))) - \nu(m(\widetilde{m}_i(w)))\right| \le C_\nu \left|m(\widetilde{m}_i(z)) - m(\widetilde{m}_i(w))\right| \le C_\nu \frac{1}{(1-r_R)^2}\,|z-w|\,,$$

which is independent of $m, \widetilde{m}_i$. Combining these with eq. (A.2-A.3) we get

$$\left|\int_{\Omega_0} f(z) d\mathrm{vol}_H(z) - \min_{m(z_i)=w_j} \sum_k \alpha_k f(p_k)\right| \le (C_\mu + C_\nu)\,\frac{\int_{\Omega_0} d\mathrm{vol}_H}{(1-r_R)^2}\varphi\left(\{p_k\}\right),$$

which finishes the proof. □

APPENDIX B.

In this appendix we provide a short exposition on discrete and conjugate discrete harmonic functions on triangular meshes as presented in [6, 23, 24, 25], and we show how this theory can be used in our context to conformally flatten disk-type (or even just simply connected) triangular meshes.

We will use the same notations as in Section 4. Discrete harmonic functions are defined using a variational principle in the space of continuous piecewise linear functions defined over the mesh $PL_{\mathcal{M}}$ ([6]), as follows. Let us denote by $\phi_i(z)$, $i = 1, .., m$, the scalar functions that satisfy $\phi_j(v_i) = \delta_{i,j}$ and are linear on each triangle $f_{i,j,k} \in F$. Then, the (linear) space of continuous piecewise-linear function on M can be written in this basis:

$$PL_M = \left\{ \sum_{i=1}^m u_i \phi_i(z) \;\mid\; (u_1, ..., u_m)^T \in \mathbb{R}^m \right\}.$$

Next, the following quadratic form is defined over PL_M:

$$E_{Dir}(u) = \sum_{f\in F} \int_f \langle \nabla u, \nabla u \rangle \, d\mathrm{vol}_{\mathbb{R}^3}, \tag{B.1}$$

where $\langle\cdot\rangle = \langle\cdot\rangle_{\mathbb{R}^3}$ denotes the inner-product induced by the ambient Euclidean space, and $d\mathrm{vol}_{\mathbb{R}^3}$ is the induced volume element on f. This quadratic functional, the *Dirichlet energy*, can be written in coordinates of the basis defined earlier as follows:

$$E_{Dir}\left(\sum_i u_i\phi_i\right) = \sum_{i,j=1}^m u_i u_j \left[\sum_{f\in F}\int_f \langle \nabla\phi_i, \nabla\phi_j\rangle\right] d\mathrm{vol}_{\mathbb{R}^3} = \sum_{i,j=1}^m u_i u_j \int_M \langle \nabla\phi_i, \nabla\phi_j\rangle \, d\mathrm{vol}_{\mathbb{R}^3}. \tag{B.2}$$

The discrete harmonic functions are then defined as the functions $u \in PL_M$ that are critical for $E_{Dir}(u)$, subject to some constraints on the boundary of M. The linear equations for discrete harmonic function $u \in PL_M$ are derived by partial derivatives of E_{Dir}, (B.2) w.r.t. $u_i, i = 1, .., m$:

$$\frac{\partial E_{Dir}(u)}{\partial u_k} = 2\sum_{i=1}^m u_i \left[\sum_{f\in F}\int_f \langle \nabla\phi_i, \nabla\phi_k\rangle\right] d\mathrm{vol}_{\mathbb{R}^3} = 2\int_M \langle \nabla u, \nabla\phi_k\rangle \, d\mathrm{vol}_{\mathbb{R}^3} = 2\int_{R_k} \langle \nabla u, \nabla\phi_k\rangle \, d\mathrm{vol}_{\mathbb{R}^3}, \tag{B.3}$$

where $R_k \subset M$ is the 1-ring neighborhood of vertex v_k. The last equality uses that ϕ_k is supported on R_k.

Now, let $u = \sum_i u_i\phi_i$ be a discrete harmonic function. Pinkall and Polthier observed that conjugating the piecewise-constant gradient field ∇u (constant on each triangle $f \in F$), i.e. rotating the gradient ∇u in each

triangle f by $\pi/2$ in the positive (= counterclockwise) sense (we assume M is orientable), results in a new vector field $*du = Jdu$ with the special property that its integrals along (closed) paths that cross edges only at their mid-points are systematically zero (see for example [25]). This means in particular that we can define a piecewise linear function $*u$ such that its gradient satisfies $d * u = *du$ and that is furthermore continuous through the mid-edges $\mathbf{v} \in \mathbf{V}$. The space of piecewise-linear functions on meshes that are continuous through the mid-edges is well-known in the finite-element literature, where it is called $ncPL_M$, the space of non-conforming finite elements [3]. The Dirichlet form (B.1) is defined over the space of non-conforming elements $ncPL_M$ as well; the non-conforming discrete harmonic functions are defined to be the functions $v \in ncPL_M$ that are critical for E_{Dir} and that satisfy some constraints on the mid-edges of the boundary of the mesh. Polthier [25] shows that if $u \in PL_M$ is a discrete harmonic function, then $*u \in ncPL_M$ is also discrete harmonic, with the same Dirichlet energy, and vise-versa. Solving for the discrete harmonic function after fixing values at the boundaries amounts to solving a sparse linear system which is explicitly given in [25].

This theory can be used to define discrete conformal mappings, and used to flatten a mesh in a "discrete conformal" manner, as follows. The flattening is done by constructing a pair of conjugate piecewise linear functions $(u, *u)$ where $u \in PL_M$, $*u \in ncPL_M$, and the flattening map $\Phi : \mathbf{M} \to \mathbb{C}$ is given by

$$\Phi = u + \mathbf{i} * u. \tag{B.4}$$

Since $d * u = Jdu$, Φ is a similarity transformation on each triangle $f \in F$. Furthermore, Φ is continuous through the mid-edges $\mathbf{v}_r \in \mathbf{V}$. This means that Φ is well-defined on the mid-edges $\mathbf{V}$ and maps them to the complex plane.

The function u is defined by choosing an arbitrary triangle $f_{out} \in F$, excising it from the mesh, setting the values of u at two of f_{out}'s vertices u_{i_1}, u_{i_2} to 0 and 1, respectively, and then solving for the discrete harmonic u that satisfies these constraints. See for example Figure 3 (top-left); the "missing mid-edge face" corresponding to the excised face f_{out} would have connected the three mid-edge vertices that have a only one mid-edge face touching them. The conjugate function $*u$ is constructed by a simple conjugation (and integration) process as described in [25] and [18].

A surprising property of the Discrete Uniformization Φ as it is defined above, which nicely imitates the continuous theory (see [29]) is that it takes the boundaries of $\mathcal{M}$ to horizontal slits, see Figure 3, top row (boundary vertices colored in red). This property allows us to easily construct a closed form analytic map (with "analytic" in its standard complex analytic sense) that will further bijectively map the entire complex plane $\mathbb{C}$ minus the slit to the open unit disk, completing our Uniformization procedure.

This property is proved by arguments similar to those for Proposition 35 in [25]; see also [18]. More precisely, we have

Theorem B.1. *Let $\Phi : \boldsymbol{M} \to \mathbb{C}$ be the flattening map from the mid-edge mesh $\boldsymbol{M}$ of a mesh $\mathcal{M}$ with boundary, using a discrete harmonic and conjugate harmonic pair as described above. Then, for each connected component of the boundary of $\mathcal{M}$, the mid-edge vertices of boundary edges are all mapped onto one line segment parallel to the real axis.*

Proof. Suppose $u = \sum_i u_i \phi_i(\cdot)$ is a discrete harmonic, piecewise linear and continuous function, defined at each vertex $v_i \in V$, excluding the two vertices of the excised triangle for which values are prescribed; then we have, by (B.3),

$$\int_{R_i} \langle \nabla \phi_i, \nabla u \rangle d\mathrm{vol}_{\mathbb{R}^3} = 0, \tag{B.5}$$

Next, consider a boundary vertex v_j of the mesh $\mathcal{M}$. Denote by $\mathbf{v}_r, \mathbf{v}_s$ the mid-edge vertices on the two boundary edges touching vertex v_j. We will show that $*u(\mathbf{v}_r) = *u(\mathbf{v}_s)$; this will imply the theorem, since $*u$ gives the imaginary coordinate for the images of the mid-edge vertices under the flattening map (see (B.4)) .

Observe that on the triangle $f_{i,j,k}$,

$$\nabla\phi_j = \frac{J(v_i - v_k)}{2\,\mathrm{vol}_{\mathbb{R}^3}(f_{i,j,k})}. \tag{B.6}$$

Recalling that $\nabla * u = J\nabla u$, using (B.6), and $J^T = -J$, we obtain

$$\begin{aligned}
*u(\mathbf{v}_r) - *u(\mathbf{v}_s) &= \int_\gamma d * u = \int_\gamma *du = \sum_{f_{i,j,k}\ni v_j} \left\langle J\nabla u\,|_{f_{i,j,k}}, \frac{1}{2}(v_i - v_k) \right\rangle \\
&= \sum_{f_{i,j,k}\ni v_j} \left\langle \nabla u\,|_{f_{i,j,k}}, \frac{1}{2}J^T(v_i - v_k) \right\rangle \\
&= \sum_{f_{i,j,k}\ni v_j} \left\langle \nabla u\,|_{f_{i,j,k}}, -\nabla\phi_j\,|_f \right\rangle \mathrm{vol}_{\mathbb{R}^3}(f) \\
&= -\int_M \langle \nabla u, \nabla\phi_j \rangle\, d\mathrm{vol}_{\mathbb{R}^3} \\
&= 0,
\end{aligned}$$

where γ is the piecewise linear path starting at $\mathbf{v}_r$ and passing through the mid-edge vertices of the 1-ring neighborhood of v_j ending at $\mathbf{v}_s$. The last equality is due to (B.3). □

A natural question, when dealing with any type of finite-element approximation, concerns convergence as the mesh is refined: convergence in what sense, and at what rate? For discrete harmonic functions over meshes, this convergence is discussed in [12, 24]. Note that these convergence results are in the weak sense; this motivated our defining the discrete conformal factors $\boldsymbol{\mu}_\mathbf{f}$ via integrated quantities (volumes) in Section 4.

Finally, we note that the method presented here for Discrete Uniformization is just one option among several; other authors have suggested other techniques; for example [10]. Typically, this part of the complete algorithm described in this paper could be viewed as a "black box": the remainder of the algorithm would not change if one method of Discrete Uniformization is replaced by another.

APPENDIX C.

In this Appendix we prove a lemma used in the proof of Theorem 4.4.

Lemma *The $N \times N$ matrices π satisfying*

$$\begin{cases} \sum_i \pi_{ij} \le 1 \\ \sum_j \pi_{ij} \le 1 \\ \pi_{ij} \ge 0 \\ \sum_{i,j} \pi_{ij} = M < N \end{cases} \tag{C.1}$$

constitute a convex polytope $\mathcal{P}$ of which the extremal points are exactly those π that satisfy all these constraints, and that have all entries equal to either 0 or 1.

Remark. Note that the matrices $\pi \in \mathcal{P}$ with all entries in $\{0, 1\}$ have exactly M entries equal to 1, and all other entries equal to zero; if one removes from these matrices all rows and columns that consist of only zeros, what remains is a $M \times M$ permutation matrix.

Proof. $\mathcal{P}$ can be considered as a subset of $\mathbb{R}^{N^2}$, with all entries nonnegative, summing to M. The two inequalities in (C.1) imply that the entries of any $\pi \in \mathcal{P}$ are bounded by 1. These inequalities can also be rewritten as the constraint that every entry of $A\mathcal{P} - b \in \mathbb{R}^{2N}$ is non positive, where A is a $\mathbb{R}^{2N} \times \mathbb{R}^{N^2}$ matrix, and b is a vector in $\mathbb{R}^{2N}$. It follows that $\mathcal{P}$ is a (bounded) convex polytope in $\mathbb{R}^{N^2}$.

If $\pi \in \mathcal{P} \subset \mathbb{R}^{N^2}$ has entries equal to only 0 or 1, then π must be an extremal point of $\mathcal{P}$ by the following argument. If $\pi_\ell = 1$, and π is a nontrivial convex combination of π^1 and π^2 in $\mathcal{P}$, then

$$\pi = \lambda\,\pi^1 + (1-\lambda)\,\pi^2 \text{ with } \lambda \in (0,1) \Longrightarrow 1 = \lambda\,\pi^1_\ell + (1-\lambda)\,\pi^2_\ell \text{ with } \pi^1_\ell\,,\ \pi^2_\ell \geq 0 \Longrightarrow \pi^1_\ell = \pi^2_\ell = 1\ .$$

A similar argument can be applied for the entries of π that are 0. It follows that we must have $\pi^1 = \pi = \pi^2$, proving that π is extremal.

It remains thus to prove only that $\mathcal{P}$ has no other extremal points. To achieve this, it suffices to prove that the extremal points of $\mathcal{P}$ are all integer vectors, i.e. vectors all entries of which are integers – once this is established, the Lemma is proved, since the only integer vectors in $\mathcal{P}$ are those with all entries in $\{0,1\}$.

To prove that the extremal points of $\mathcal{P}$ are all integer vectors, we invoke the Hoffman-Kruskal theorem (see [15], Theorem 7C.1), which states that, given a $L \times K$ matrix $\mathbb{M}$, with all entries in $\{-1,0,1\}$, and a vector $b \in \mathbb{R}^L$ with integer entries, the vertices of the polytope defined by $\{x \in \mathbb{R}^K\,;\, (\mathbb{M}x)_\ell \leq b_\ell \text{ for } \ell = 1,\ldots,L\}$ are all integer vectors in $\mathbb{R}^K$ if and only if the matrix $\mathbb{M}$ is totally unimodular, i.e. if and only if every square submatrix of $\mathbb{M}$ has determinant 1, 0 or -1.

We first note that (C.1) can indeed be written in this special form. The equality $\sum_{i,j} \pi_{ij} = M$ can be recast as the two inequalities $\sum_{i,j} \pi_{ij} \leq M$ and $-\sum_{i,j} \pi_{ij} \leq -M$. The full system (C.1) can then be written as $(\mathbb{M}\pi)_\ell \leq b_\ell$ for $\ell = 1,\ldots,L$, where $\mathbb{M}$ is a $(2N + 2 + N^2) \times N^2$ matrix constructed as follows. Its first $2N$ rows correspond to the constraints on the sums over rows and columns; the entries of the next row are all 1, and of the row after that, all -1 – these two rows correspond to the constraint $\sum_{i,j} \pi_{ij} = M$; the final $N^2 \times N^2$ block is diagonal, with all its diagonal entries equal to -1. The first $2N$ entries of b are 1; the next 2 entries are M and $-M$; its final N^2 entries are 0. By the Hoffman-Kruskal theorem it suffices thus to show that $\mathbb{M}$ is totally unimodular.

Because the last N^2 rows, the *bottom rows* of $\mathbb{M}$, have only one non-zero entry, which equals -1, we can disregard them. Indeed, if we take a square submatrix of $\mathbb{M}$ that includes (part of) one of these bottom rows, then the determinant of the submatrix is 0 if only zero entries of the bottom row ended up in the submatrix; if the one -1 entry of the bottom row is an entry in the submatrix, then the determinant is, possibly up to a sign change, the same as if that row and the column of the -1 entry are removed. By this argument, we can remove all the rows of the submatrix partaking of the bottom rows of $\mathbb{M}$.

We thus have to check unimodularity only for $\mathbb{M}'$, the submatrix of $\mathbb{M}$ given by its first $2N + 2$ rows. If any submatrix contains (parts of) both the $(2N + 1)$st and the $(2N + 2)$nd row, then the determinant is automatically zero, since the second of these two rows equals the first one, multiplied by -1. This reduces the problem to checking that $\mathbb{M}''$, the submatrix of $\mathbb{M}$ given by its first $2N + 1$ rows, is totally unimodular.

We now examine the top $2N$ rows of $\mathbb{M}''$ more closely. A little scrutiny reveals that it is, in fact, the adjacency matrix $\mathbb{G}$ of the complete bipartite graph with N vertices in each part.[1] It is well-known (see e.g. Theorem 8.3 in [28]) that this adjacency matrix is totally unimodular, so any square submatrix of $\mathbb{M}''$ that does not involve the $(2N + 1)$st row of $\mathbb{M}''$ is already known to have determinant 0, 1 or -1. We thus have to check only submatrices that involve the last row, i.e. matrices that consist of a $(n-1) \times n$ submatrix of $\mathbb{G}$, with an added nth row with all entries equal to 1. We'll denote such submatrices by $\mathbb{G}'$.

We can then use a simple induction argument on n to finish the proof. The case $n = 2$ is trivial. In proving the induction step for $n = m$, we can assume that each of the top $m - 1$ rows of our $m \times m$ submatrix $\mathbb{G}'$ contains at least two entries equal to 1, since otherwise the determinant of $\mathbb{G}'$ would automatically be 0, 1 or -1 by induction.

[1] The adjacency matrix A for a graph $\mathcal{G}$ has as many columns as $\mathcal{G}$ has edges, and as many rows as $\mathcal{G}$ has vertices; if we label the rows and columns of A accordingly, then $A_{ve} = 1$ if the vertex v is an end point of the edge e; otherwise $A_{ve} = 0$. An adjacency matrix thus has exactly two nonzero entries (both equal to 1) in each column. The number of nonzero entries in the row with index v is the degree of v in the graph.

The first $m-1$ rows of $\mathbb{G}'$ correspond to vertices in the bipartite graph, and can thus be partitioned into two sets S_1 and S_2, based on which of the two parts of N vertices in the graph they pertain to. Let us call S the larger of S_1 and S_2; S consists of at least $\lceil \frac{m-1}{2} \rceil$ rows. Let us examine the $(\#S) \times m$ sub-matrix $\mathbb{G}''$ constructed from exactly these rows. We know that each column of $\mathbb{G}''$ has exactly one entry 1, since all the rows of $\mathbb{G}''$ correspond to the same group of vertices in the bipartite graph. Therefore, summing all the rows of $\mathbb{G}''$ gives a vector v of only zeros and ones; since each row in $\mathbb{G}''$ contains at least two entries equal to 1, the sum of all entries in v is at least $2\left(\lceil \frac{m-1}{2} \rceil\right) \geq m-1$. The vector v has thus at least $m-1$ entries equal to 1; the remaining mth entry of this linear combination of the top $m-1$ rows of $\mathbb{G}'$ is either 1 or 0. In the first case, the determinant of $\mathbb{G}'$ vanishes, since its last row also consists of only ones. In the second case, we can subtract v from the last row of $\mathbb{G}'$ without changing the value of the determinant; the resulting last row has all entries but one equal to 0, with a remaining entry equal to 1. The determinant is then given by the minor of this remaining entry, and is thus 0, 1 or -1 by the unimodularity of $\mathbb{G}$. □

Princeton University